UNIVERSITETET
I OSLO

Masteroppgave

How is it operationalized

How examiners operationalize the assessment criteria given during an English examination in lower secondary school in Norway

Tollef Kristensen

Engelsk fagdidaktikk
30 studiepoeng

Institutt for lærerutdanning og skoleforskning
Det utdanningsvitenskaplige fakultet

How is it operationalized?

How examiners operationalize the assessment criteria given during an English examination in lower secondary school in Norway

Mastergradsavhandling i engelsk fagdidaktikk

Tollef Kristensen

UNIVERSITETET I OSLO

Våren 2023

2023

«How is it operationalized»

Tollef Kristensen

https://www.duo.uio.no

Abstract

The present thesis investigates how examiners operationalize the assessment criteria they are given during an oral examination in English. Previous research has shown that there is variation in what assessment criteria examiners focus on during assessment. This thesis focuses on how these criteria are discussed by the examiners, as knowledge of how they create a shared understanding of said criteria is beneficial to the reliability and validity of assessments.

This is a qualitative study, using audio recordings of examinations which were gathered by the ETOS project. Two teachers worked as examiners for five different students in a lower secondary school. The recordings included their conversations during the students' examinations and their conversations afterward where they discussed different aspects of the student's performance. The data were analysed using a combination of inductive and deductive approaches, with the assessment criteria forming the basis for coding.

The findings suggest that the examiners generally agree on how to interpret the different assessment criteria. However, it also shows that examiners on average focused more on the assessment criteria related to content than on those related to language. They also took a holistic approach to some criteria, like fluency, while focusing on more specific and easily observable factors for other criteria, like vocabulary. Language ideology was not mentioned by examiners, despite the influence it might have on their understanding of criteria like pronunciation. Finally, the examiners discussed aspects of the student's performance not mentioned in the criteria.

The implications of this thesis include suggestions that more rater training is conducted to ensure a shared understanding of the assessment criteria, possibly through the use of audio recordings. Teachers should be made aware of their own biases related to topics like intelligibility and nativeness, and how this might influence their assessment of a student's pronunciation.

Sammendrag

Denne masterstudien undersøker hvordan sensorer operasjonaliserer vurderingskriteriene de får under en muntlig eksamen i engelsk. Tidligere forskning har vist at det det er variasjon i hvilke vurderingskriterier sensor fokuserer på i løpet av en eksamensvurdering. Derfor ser denne studien på hvordan disse kriteriene blir diskutert av sensorene, ettersom kunnskap om hvordan de skaper en felles forståelse av disse kriteriene kan være nyttig for å sikre reliabilitet og validitet i vurderinger.

Dette er en kvalitativ studie, som har brukt lydopptak som var samlet inn av ETOS prosjektet. To lærere var sensorer for fem forskjellige studenter i en ungdomsskole. Disse lydopptakene inkluderte samtalen deres under selve eksamen, samt samtalen deres etterpå hvor de diskuterte ulike deler av studentens framtreden. Disse dataene ble analysert ved å bruke en kombinasjon av induktive og deduktive tilnærminger, hvor vurderingskriteriene formet et fundament for kodingen.

Funnene viste at sensorene generelt var enige om hvordan vurderingskriteriene skulle tolkes. Men det kom også fram at sensorene i gjennomsnitt fokuserte mer på vurderingskriteriene som var relatert til innhold enn de som var relatert til språk. De tok også en holistisk fremgangsmåte til noen kriterier, som flyt, mens de fokuserte på mer spesifikke og lett observerbare faktorer for andre kriterier, som vokabular. Språkideologi ble ikke nevnt, selv om dette kan påvirke kriterier som uttale. Sensorene diskuterte deler av studentens framtreden som ikke var nevnt i vurderingskriteriene.

Implikasjonene av denne masteroppgaven er blant annet at det gjennomføres mer trening for sensorer for å sikre at de har en delt forståelse av vurderingskriteriene. Lydopptak kan muligens hjelpe. Lærere burde også gjøres mer klar over deres egne subjektive forståelser, eksempelvis knyttet til om forståelighet eller å høres ut som en førstespråkstaler burde være målet, og hvordan dette kan påvirke deres tolkning av en elev sin uttale.

Acknowledgments

First and foremost, I would like to thank my supervisor Ulrikke Rindal, whose useful comments and feedback throughout the writing of this study have been indispensable. I will forever be grateful for the time you have spent giving me advice on how to further improve my thesis. Without your dedicated support, this thesis would have looked very different. I would also like to show my deep appreciation to Lisbeth Myklebostad Brevik, co-supervisor of my thesis and project leader for the study whose data I used. Your feedback, particularly on the topic of methodology, doubtlessly raised the quality of this thesis and helped to encourage me to do my best. Finally, my fellow English Didactics students deserve praise as well, for it might take a village to raise a child but it takes a community to write a thesis. Thank you for making that process smoother than it would have otherwise been.

Writing this thesis was by no means an easy task, and I underestimated the workload it entailed in the beginning. There have certainly been many ups and downs throughout the process. I would be remiss if I did not mention the people who made it possible for me to get through it all. I would not have endured without the love and support given to me by my family throughout it all. Thank you, Tom, Celine, and my little brother Tobias. I love you, I always will, and thank you for always believing in me and for supporting me no matter what. It means more to me than can ever be put into words. Finally, thank you to all my friends who have endured me rambling about my thesis one day, only to ramble about another aspect the next day. Truly, Atlas himself would nod in appreciation of your patience. You are the best friends anyone could ever hope for.

این نیز بگذرد,

This too shall pass.

Tollef Kristensen

Blindern, May 2023

Table of contents

1.0 Introduction

In this thesis, I will discuss the complexity of assessing oral proficiency in Norway and the implications this has for oral English examinations in a Norwegian lower secondary school. Throughout my time studying to become a teacher, oral exams have always interested me. Oral exams are the final assessment after a year of learning, where the examiners must discuss and reach a shared understanding in a very limited amount of time. From the perspective of a student, the examination appears to be a black box. You provide input in the form of your performance, and then receive output in the form of a grade. The process that examiners utilize to turn the students' performance into something that can be assessed and graded was unknown and thus very interesting to me.

The national curriculum says what students are expected to know by the end of a term, and thus constrain what teachers can do (Eriksen, 2018). Yet it is also contextualized, as the oral exams are locally developed. Teachers have a high degree of autonomy in what to teach and assess, though this varies from school to school and can be limited by various approaches such as how much emphasis the school puts on teacher cooperation (Eriksen, 2018). While classroom assessment has both a formative and a summative character, an examination is intended to be summative. It is meant to assess what knowledge and skills the student can display in that exact time and space. This assessed level of competence is then used to assign a grade to the student that is appropriate for that level of achievement. Bachman and Palmer (2010) say that high-stakes decisions are those that can have significant consequences for the lives of individuals, and because the grades given at an end-of-the-term examination will impact the students' future studies, it can be argued that exams are high-stakes.

Operationalization is the process of turning criteria into something that can be used for assessment. Studies suggest that teachers find it difficult to operationalize the competence aims in the national curriculum into concrete learning objectives, and that "[t]here does not seem to be a shared understanding of what constitutes adequate, good, and excellent performance in different subject areas" (Nusche et al., 2011, p. 129). This is potentially problematic, as a key aspect of the examination is the grade the student receives at the end.

The assessment criteria are operationalizations of the competence aims which are set by the Directorate of Education. However, these assessment criteria must then be operationalized by the examiners. Therefore, this study intend to look into how examiners work to understand the descriptors in the assessment criteria they are given during an oral examinations.

1.1 Research Aim and Purpose

Research specifically regarding the assessment of oral English at the upper secondary level in Norway is found in a doctoral thesis by Henrik Bøhn (2016). He found that teachers generally agree on the grading of students' oral English exam performances, but that their views differed regarding which performance aspects they should focus on. When teachers did disagree, it was often about narrow performance features.

Teachers were also reported to hold different opinions on the relationship between the assessment of language and content, and if one were more important than the other. Some teachers also assessed features of student performances not relevant to the competence aims in the English subject curriculum, such as effort (Bøhn, 2016). Since the guidelines teachers use for assessment are quite general, teachers may develop their understanding based on their interpretation of assessment criteria and competence aims (Rindal, 2015). Because this process is unique to each teacher, the examiners must reach a shared understanding during the examination, to ensure that they are talking about the same things.

The assessment criteria used during the examinations had a combination of nouns to define what competence is being assessed, verbs to define how the student is to show that competence, and adjectives to define what separates each level of achievement. These words, particularly the adjectives used to help narrow down what level the performance was at, must be understood by the examiners to define the student's level of achievement. Examples of these sentence structures are "has mostly correct pronunciation" and "has very good pronunciation". Each examiner must interpret these words on their own through a process of operationalization, to turn them into something that can be measured and used for assessment.

This thesis is a response to Bøhn's (2016) call for further research on the assessment of oral English. The main way examiners can reach a shared understanding during examinations is

through discussions of the criteria and their observations during the examinations, and this thesis aims to investigate these discussions to see how examiners share their understandings.

1.2 Operationalization of assessment criteria

The operationalization of assessment criteria is left to the local level (UDIR, 2013a), which means that it is up to the individual teachers to reach an understanding of them. It is not clear what arguments they bring up, what examples they use, and how these examples are discussed by the examiners. This thesis intends to investigate this black box. Assessment depends on the collection of evidence of the student's knowledge, skills, and abilities, which are then used by the examiners to decide what level of achievement the student is at (Green, 2014). The use of two examiners during the exam influences this, as two people are likely to notice a larger and more diverse amount of evidence of the student's knowledge, but they also need to communicate this to each other.

An example of this is the construct of fluency. Fluency can be operationalized using indicators like speech rate or pausing, which can be clearly defined and measured (Luoma, 2004). Speech rate can, theoretically, be determined by timing and counting the number of words the student says, while pausing can be found if the examiners count the number of pauses. While this is possible, it is not practical within the confines of an oral examination, and it is also debatable if this level of detail is useful to the examiners (Bøhn, 2016). It is therefore interesting to see how examiners attempt to clarify such criteria.

1.3 English didactics

The expanded role that English has obtained thanks to globalization means that it is a very useful language to know. Most Norwegians are exposed to it in their daily lives thanks to audio, visual media, and written texts (Rindal, 2014). English is viewed as important by the Norwegian Directorate for Education and Training because of its role as a tool and a means of communication with other people (UDIR, 2013b). While it is often regarded as a foreign language, Rindal (2015, 2020) argues that English has many of the characteristics of a second language.

English has some characteristics that teachers can interpret in different ways and thus disagree about, without necessarily being aware of it. Because there are no guidelines saying which interpretation is correct, it is left up to the individual examiner, which makes it important for them to discuss this. Whether they are aware of it or if its effects are mostly subconscious, the examiners' language ideology can impact how they assess a student's understanding. Some examiners might believe that the use of nonstandard variants of English helps with communication while others believe this merely creates additional confusion (Iannuzzi & Rindal, 2018). If one teacher believes that additional languages are a resource and another teacher views it as a deviation from how English should be spoken, this will influence their view of the students' language.

1.4 Research question

The overarching aim of this thesis is to discuss the assessment of oral proficiency, what makes the Norwegian assessment situation complex, and the challenges which this leads to for individual examiners. As part of this exploration, I will identify what teachers pay attention to when it comes to oral assessment. In addition, I explore how teachers develop a shared understanding of the assessment criteria they are given before the examination. The following research question will serve to guide this thesis:

How do examiners operationalize the assessment criteria used during an oral examination in English in lower secondary school?

Because the purpose of this study is to gain insight into how the examiners operationalize the assessment criteria, I concluded that it would be important to gain insight into what examiners did during an examination. Rather than going for an overview of many teachers, I wanted to look in depth at what some examiners did, and a qualitative study was the type most suited for this goal. Interviews could gain insight into the examiners' thoughts before or after an examination, but observation would allow me to learn what took place during the assessment situation itself.

1.5 The ETOS project

I was fortunate enough to be invited by the ETOS project leader Lisbeth M. Brevik, Professor at the Department of Teacher Education and School Research at the University of Oslo, to use

the data material from the ETOS project for my thesis. The ETOS project looked into two lower secondary schools that offered bilingual teaching (Brevik & Doetjes, 2020). The ETOS project's information page at the University of Oslo describes its main aim thusly:

The ETOS project aims to increase our knowledge of bilingual education, which is instructed partly in Norwegian and partly in English. ETOS will consider student motivation, learning outcomes, and perceived relevance across individual subjects [...] The evaluation considers both language and content aspects of the instruction.

1.6 Thesis structure

The present chapter has contextualized the study by providing a general introduction and background information. Chapter 2 explains the theoretical framework of the thesis, largely based on theories from the fields of educational assessment, language and content as well as teachers' cognition within the framework of English as a subject. These theories provide relevant conceptualizations for understanding how teachers operationalize the assessment criteria that are used during the examinations. It also shows how non-criterion-based aspects of the student's performance can be brought up by the examiners during their discussions. In addition, what is being assessed by the examiners will be explored by looking at the English subject curriculum, the Common European Framework of Reference for Languages, and central aspects of oral competence. Lastly, assessment theory connected to oral competence will be accounted for. Chapter 3 outlines the research design, including the research method, the research question, the participants, the data collection, and the framework used for the analysis. Possible limitations and ethical considerations regarding empirical research are also discussed in this chapter. In Chapter 4, the findings from the analysis of the data material are presented. Chapter 5 will outline the main findings and then discuss them in light of theory as well as previous research where applicable. Finally, my concluding remarks and suggestions for future research are presented in Chapter 6.

2.0 Theory

In this chapter, I discuss the theoretical framework of the thesis. This thesis aims to explore how teachers operationalize the assessment criteria when they are assessing a student's oral performance alongside another teacher. The assessment of a student's exam involves many factors, which might have implications for what teachers understand as important to assess. Teachers working in Norwegian schools have much autonomy. They must interpret the national curriculum and the subject curriculum, and then decide what and how they are going to teach and assess. This introduces a wide spectrum of possible interpretations and understandings. The national curriculum that will be used is LK06, as it was the curriculum in use in lower secondary school at the time.

I will introduce relevant theories tied to the nature of examinations (2.1), including the differences between summative and formative assessment, formal and informal assessment, and norm-referenced and criterion-referenced assessment. An overview of particularly relevant aspects of oral assessment theory and the nature of speaking will be outlined (2.2). Theoretical perspectives on the two common aspects of assessment, content and language, will be reviewed (2.3). The importance of reliability (2.4) and validity (2.5) in the assessment setting will be accounted for. Then, the role of equity when it comes to assessment will be explored. Finally, an account of previous assessment research (2.8) in an international context (2.8.1) and the context of Norwegian education (2.8.2) will be provided.

2.1 Assessment theory for examinations

Here I will review relevant theories to discuss the structure of the examination and the role it plays in the student's education within the Norwegian educational framework. The reason for this review is that the examination's structure might influence how the examiners and the examinee act during the examination.

2.1.1 Summative and formative assessment

According to the regulation relating to the Education Act, the Norwegian assessment system is based on individual assessment (Forskrift til opplæringslova, 2006, § 3-22). This assessment can take two forms, either formative or summative assessment. Formative assessment is done throughout the learning process, while the students are in the classroom. Meanwhile,

summative assessment is testing that is done at the end of the learning process. This summative assessment can be written, oral or practical. Summative assessment is connected to the competence aims of the subject curriculum. It is intended to show the level of competence the students have in the topics specified in the curriculum at a specific moment in time. Based on this, the English oral examination at the lower secondary level in Norway can be defined as a form of summative assessment.

2.1.2 Formal and informal assessment

Simensen (1998) distinguishes between informal and formal assessment in education. Informal assessment is the assessment the teacher does daily, such as dialogues with students or observations of classroom activities. Formal assessment, on the other hand, is done through the use of examinations and other tests. This means that all tests given during the school year are part of the formal assessment. Based on this, the exams in lower secondary school can be said to be a formal assessment situation.

2.1.3 Norm-referenced and criterion-referenced assessment

Brown (1996) says that a norm-referenced test is intended to measure global language abilities, which are performance features like language proficiency or reading comprehension. The score one student receives is compared with the scores of all the other students who took the test. The results of the test are then spread out in a distribution curve, where the students who scored lowest compared to their peers receive low grades, and those who scored high on the test are given high grades. Thus, in a norm referenced test, the grade which the students are given is dependent on how other students performed during the test.

Meanwhile, a criterion-referenced test measures well-defined and specific objectives, which are specific to what is being assessed. A student's score shows how much the student has learned of the objectives that are tested, based on criteria that are used for all of the students. If all students know all the objectives well, they would all get the highest grades (Brown, 1996). Thus, a criterion-referenced test is less influenced by the surrounding context than a norm-referenced test is.

If we go to the Regulations of the Education Act, we see that the regulations specify that assessment should be criterion-referenced in Norway (Forskrift til opplæringslova, 2006, § 3-

22). This means that the students are to be assessed according to the competence aims of the English subject curriculum and that they should not be compared to each other. The examination is intended to test the degree to which the student has understood the material they have studied and give them a grade based on this. Thus, the lower secondary English oral examination in Norway can be said to be a summative and formal criterion-referenced test.

2.2 Oral assessment theory

In this section, I will look at what characterizes the assessment of oral production, with a particular focus on what it is important to pay attention to when one assesses speaking, as certain parts of oral communication are particularly difficult to assess.

2.2.1 Oral assessment in general

The assessment of oral production is made even more difficult than other types of assessment due to the nature of speaking in itself (Luoma, 2004). Luoma says that it is especially challenging to assess speaking because there are so many different factors that influence the way we evaluate oral proficiency (2004). Because the examiner needs to keep many factors in mind at the same time and only has a limited amount of time to sort these thoughts, this can cause difficulty in processing all the relevant information. The competence aims looks at two different aspects of oral competence, the ability to listen and the ability to speak. These aspects of oral competence can be operationalized in different ways based on the language learning paradigm followed by the examiner.

This means that how well the students do on the examination will depend on how well they have understood the tasks given as well as how well the examiner thinks they answered based on the examiner's understanding of the task. In the following paragraphs, I will look at the nature of speaking, to elaborate on why Luoma says that the assessment of speaking is so challenging.

2.2.2 The nature of speaking

Elements that are typically considered to be important aspects of oral proficiency include accent, grammar, vocabulary, what mistakes and errors are made, and the ability to use language appropriate to context the student is in (Luoma, 2004). Accent, or the sound of speech,

is difficult to assess because it is not easy to define what is correct speech. Some claim that correct speech is how native speakers of a language speak, while others argue that intelligibility is what the teacher should focus on. (Levis, 2005) There is no mention of a native speaker model in the English subject curriculum, and the same can be said for intelligibility (UDIR, 2013b). This makes it difficult to agree on a standard that students should be assessed according to, as it is something that can differ greatly between examiners (Luoma, 2004). Speaking can be said to be the most difficult skill to assess reliably because of this lack of a shared standard.

Even though students can achieve functional and understandable speech, it is very difficult for them to reach the level of a native speaker. This means that most would receive lower grades if they were to be assessed according to the standard of native speakers (Luoma, 2004). When it come to the sound of the students' speech, there are two elements which the examiner can focus on; the accuracy of pronunciation, and the expressiveness of the speaker's voice (Luoma, 2004). It is often tempting to focus on pronunciation because that can be measured against a standard, even if it can be difficult to choose said standard, as previously noted (Luoma, 2004). Regardless of whether they choose to focus on intelligibility or nativeness, or even both, the examiners must be conscious of what they want to focus on and make this clear to each other, as this can influence their overall opinion of the students' speech.

Luoma also notes that grammar is an element that influences the way we evaluate spoken utterances. It is important to take into account that the grammar used for speech differs from that used for writing. People usually do not speak in complete sentences, but rather in idea units (Luoma, 2004). These units are phrases that are connected with small words or with small pauses between them, also known as filler words. This makes the grammar in speech quite different from written grammar, where filler words are viewed as something to be avoided completely. Additionally, planned speech, like the students' prepared presentations, should contain more of the grammatical structures found in written grammar than unplanned speech. The same relates to the level of formality in speech. A higher level of grammar is to be expected in a formal situation than in an informal situation, and the speaker should know this and change their grammar accordingly (Luoma, 2004). This would be the case with an examination, which is a formal situation. However, when a student is asked a follow-up question or a question they haven't prepared for, their reply will be characterized by unplanned speech. It is therefore important that an examiner is conscious of these differences in spoken grammar.

The third element Luoma mentions is vocabulary, which can be defined as the choice of words in speech. She says that the description of the higher levels of vocabulary use often mentions the ability to "express oneself precisely and provide evidence of the richness of one's lexicon" (Luoma 2004, p. 16). It is, however, important to keep in mind that in spoken language, every-day and high-frequency words are by their very nature the most common terms, and managing to use these commonly occurring words appropriately when speaking can also be viewed as a sign of advanced speaking skills (Read, 2000). The use of high-frequency words is important in spoken language, unlike in written language where the use of more specific and varied words is much more common, because a speaker does not have much time to consider their word choices. High-frequency can be exemplified as words like “this”, “that one”, “that thing”, “fine” and “good”, while specific words are words that for example replace “this” and “that” with the a specific word like “car”. Spoken language has to be easier and faster because of the nature of conversations.

The use of generic words does not always come naturally to language learners, because not all of them speak English outside the classroom, where the use of generic words is much more frequent. While this has been reduced somewhat by the increased access to computers and the internet, this suggests it might be beneficial to include the use of generic words in assessment criteria, to show learners as well as examiners that the use of these words is important signifiers of the natural use of spoken language (Luoma 2004). Furthermore, the appropriate use of fillers or hesitation markers is important since they allow the speaker to create time to talk and to speak naturally and fluently.

The fourth element Luoma discusses is slips and errors that occur in speech. There will always be errors in spoken languages, such as mispronunciations or the use of incorrect words. If a listener notices that native speakers have such errors in their language, they usually excuse the speaker because they believe that they know how it is supposed to be. Meanwhile, when a second language learner has such errors, it is often considered to be caused by a lack of competence (Luoma 2004). Because of this, Luoma says that examiners should be trained in not counting all errors they hear since this is a natural part of spoken language even for native speakers.

The last element Luoma mentions is that the speaker should use language appropriate to the purpose of speaking. A speaker must manage to use language appropriate to the situation they

are in. The ability to use appropriate language is important for the student to be considered to be a fluent speaker and it depends on many factors such as the participants in a conversation, the goal of speaking and what is going to be said in the conversation (Luoma 2004).

2.3 Theoretical perspectives on content and language

According to Met (1998), language education can be understood as a continuum from language-driven approaches to content-driven approaches, and individual countries' approaches belong somewhere on this continuum. Systems that use a language-driven approach view content as a useful tool that can be used to further learning goals, but teachers and students are not held accountable for the learning outcomes when it comes to this content. A content-driven approach, meanwhile, views content as the main purpose of the teaching and the student's acquisition of the foreign language is viewed as less important.

Competence aims such as "[The student shall be able to] describe and reflect over the situation of indigenous peoples in English-speaking countries" (UDIR, 2013b) can be interpreted as showing that content is an important part of what is to be taught and tested, while competence aims like "(The student shall be able to) understand and use a general vocabulary tied to different subjects" can be understood as showing the importance of language. Consequently, the English subject taught at the lower secondary level in Norway can be said to be located somewhere around the middle of this continuum, as both language and content constructs are to be taught to students and assessed by the teacher (Bøhn, 2016)

2.3.1 Content

'Content' is one of the identified constructs in the English subject curriculum relevant to oral assessment, and may be found in the subject area of *language learning* as well as the subject area of *culture, society, and literature* (UDIR, 2013b). The former focuses on the process of language learning the ability of students to self reflect, while the latter focuses on cultural understanding of societies and the role of international English.

As Bøhn (2016) notes in his doctoral thesis, there is very limited theoretical support to be found for the analysis of content when it comes to oral assessments. Literature on language assessment has primarily paid attention to the use of language assessment to make inferences about the student's communicative language abilities, such as language knowledge and

strategic competence, as opposed to content knowledge (Snow & Katz, 2014). Bachman and Palmer (1996), in their model of communicative competence, claim that topical knowledge "Needs to be considered in a description of language use because it provides the information that enables them to use language concerning the world in which they live, and hence is involved in all language use" (p. 65). There are, however, no comments on how this topical knowledge is to be understood by teachers nor how it can be operationalized to be useful for assessment,

Meanwhile, the CEFR (Council of Europe, 2001) describes content as a combination of the students' sociocultural knowledge, intercultural awareness, and knowledge of the world, but it is not quite clear how these constructs are to be understood and used. This was a deliberate choice, however, as the CEFR does not try to define what should be taught because that is something that can differ based on the target language in question and the pedagogic culture of a country (North, 2004).

While this is a good point, the lack of a content-specific definition in the CEFR makes it rather difficult to operationalize content for use during assessments. In Norway, the English subject curriculum is intended to be flexible to allow the teachers to adapt it to local contexts, which is done by granting relative autonomy to teachers so they can choose what content they want to teach their students and then assess this to see what they retain (Eurydice, 2008). While this can be beneficial for classroom education and assessments, it can become a problem when it comes to examinations, as the other examiner have adapted to a different local context.

Theoretical support for the assessment of 'content' can be found in Anderson and Krahtwohl's (2001) revision of Bloom's Taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). This is a framework that is utilized by Bøhn (2016) in his thesis. This framework is used to classify educational objectives and was developed to help improve educational practices. The framework were intended to encourage teachers to make sure that instruction and assessment were in line with the larger educational objectives they were trying to reach (Anderson & Krathwohl, 2001) These learning objectives typically contain a noun and a verb, where the noun described what the student was supposed to have learned, while the verb described the cognitive process the student was expected to use (Anderson and Krathwohl, 2001).

This construction of learning objectives is relevant for the Norwegian school system, as our competence aims are formulated similarly. An example of this is a competence aim from *culture, society, and literature*, which makes clear that a student should be able to "Evaluate different sources and use content from the sources in an independent, critical and verifiable manner" (UDIR, 2013b).

Based on this way of constructing learning objectives, Anderson and Krathwohl (2001) developed a two-dimensional framework, where one dimension relates to *knowledge* and the other to cognitive processes. The *knowledge* dimension relates to subject matter content and is divided into four general types of knowledge. This dimension may be related to the main subject area *Culture, Society, and Literature* in the subject curriculum. The *cognitive* process dimension is divided hierarchically from simple to complex understanding, and may be related to the main subject area *Language Learning* (Anderson and Krathwohl, 2001)

2.3.2 Language

Today, the notion of communicative competence is widely accepted as a reference point in the assessment of oral language proficiency. Initially introduced as a reaction to grammar-focused theories of language (Luoma, 2004), communicative competence sees that linguistic competence alone is not enough to explain the complex nature of how languages are employed during social interactions.

While both LK06 and the CEFR agree that competence is a broader concept than just a skill or a piece of knowledge that can be obtained on its own, Bagarić and Djigunović (2007) claim that competence is a controversial term subject to differing interpretations. Hymes (1972) broadened its definition by arguing that social knowledge is equally important as it includes the unwritten rules of when to speak and how to interact with different people for different purposes. This understanding of communicative competence focused on the communicative aspect and reiterated that language must be understood as something broader than just the sounds which the speaker produces. Canale and Swain (1980) defined communicative competence as a merger between grammatical, sociolinguistic, and strategic competencies, thus giving communicative competence the role of an overall category that contains both the constructs of language and that of content. Bachman (1990) suggested the adoption of "communicative language ability" as a more appropriate term for the theory than that of

communicative competence, arguing that said theory consists of knowledge as well as the capacity to use that knowledge.

Based on the above paragraph, there seems to be a consensus that in addition to knowing a language, a learner should also be able to use this knowledge for the purpose of communication. There are, however, challenges to these theoretical models of communicative competence. Some argue that the models don't take contextual factors like participants, the task and the setting into account, as this can influence the interaction (McNamara 2003). This can be a problem for assessment as it introduces more subjective factors which will not be present in all assessment situations, and is thus worth keeping in mind.

Some also argue that the models are already too complex to be utilized in assessment situations. McNamara (2000) notes that the models fail to adequately display the way that different aspects will influence each other during speech. This fits with Chalhoub-Deville's argument that "the ability components the language user brings to the situation or context interact with situational facets to change those facets as well as to be changed by them" (2003, p. 372). Language ability influences and is influenced by the situation, and it is thus something that is constructed by the participants in each interaction. While this is interesting for oral examination based on locally administered rating scales as well as exams, Chalhoub-Deville (2003) points out that this can cause problems for assessment as it makes generalizations difficult.

Finally, it is worth discussing the perspective of LK06 (UDIR, 2013b). The national curriculum emphasizes the role of communication, and a key aspect of the competence aims is that the speaker needs to adapt to their circumstances and their interlocutors. While some English teachers might view all grammatical mistakes as equally important to focus on and correct, others might not worry about it as long as it does not hinder communication. Bentsen (2017) brings up an example of this when she notes that subject-verb concord is not important, while tense, which can impact communication, is important. Such differences in understanding of what is important and what is not can influence how advanced an examiner views a student's language as being.

2.4 Reliability of assessment

Reliability is a central concern in language assessment and can be defined as the

ideal that the measurement should deliver the same result for the same performance as consistently as possible. (Fulcher & Davidson, 2007). Sari Luoma (2004) points out that speaking is the most difficult language skill for teachers to assess reliably. The reason for this is that a student's speaking ability is to be assessed in a real-time face-to-face interaction, where the examiner must assess what is being said and how it is being said, while the student is saying it. At the same time, the examiner must also make sure to remember what has been said, while they also must be prepared to ask a question to the student based on what the student just told them. A student might perform differently during an oral presentation for a teacher they have known all year, compared to when it is done in front of this teacher in addition to an examiner they do not know. Additionally, it is also possible that they might utilize qualitatively better language if the questions are about topics they're comfortable with, while their oral competence might appear lower if the questions enter a territory they are not as comfortable talking about.

No test can ever be considered perfectly reliable given the inherent constraints of the situation. (Luoma, 2004). This being the case, a key aspect of a test is to try and ensure that the variations in scores are caused by differences that are relevant as opposed to differences caused by irrelevant factors like who does the grading, what parts of the curriculum were chosen for the test or whether or not the student had slept well before the test. (Black and Wiliam, 2012).

In Norwegian upper secondary education as well as tertiary education, grades tend to be the main tool used when it comes to making decisions about who gets admitted to what school and whether they get to study that which they want or not. The average of their grades is used to decide who gets admitted, and this admission is generally norm-referenced, which can additionally put pressure on the examiner since they know the importance of good grades for the student's future. (The Norwegian Universities and Colleges Admission Service, 2020). This makes reliability a crucial aspect of examinations and should be given serious attention. As Harlen (2012) points out, "the more weight that is given to the summative judgment, the more stringent the quality assurances need to be" (p. 97). These quality assurances generally come in the form of the standardization of rating, though it is not clear just how standardized the assessment should be.

Moss (1994) points out that standardization can lead to a lack of focus on those intellectual activities which it is impossible to reliably document, which can further lead to a focus on teaching the test rather than keeping a holistic focus on the student's development. Taylor and

Galaczi (2011) note that standardized rating scales can lead to assessment criteria that do not adequately capture all the criteria the examiners are testing the student in. This then leads to tension between reliability secured through standardization, and a broad overview more dependent on the examiners' subjective understanding of the whole.

Another key aspect of reliability is rater training, which is intended to ensure that all examiners have the same understanding of what is to be assessed, usually using real examples (Luoma, 2004). While this requires the examples themselves to be reliable, another potential problem is that no two examples will ever be the same because of the different situational factors, which means there will always be some degree of subjectivity. Thus, rater training alone cannot solve the problem.

2.5 Validity of assessment

Validity is viewed as a vital concern in language assessment, and it is usually used to refer to the quality or 'soundness' of an assessment procedure (Luoma, 2004). Newton (2012) argues that the quality of the test itself does not matter if the test is used incorrectly or if the context for which it is being applied is not that which was originally intended by the creator. Thus, context is important for validity as well as reliability.

Today, the consensus view is that validity is the property of the inferences made from the assessment results themselves (Bøhn, 2016) Two aspects of this are of particular importance when it comes to operationalization. The first, the interpretation of test scores, is also known as score meaning. This can be exemplified by how the regulations (Forskrift til opplæringslova, 2006, § 3-5) define grade 4. It is stated that the grade means the student has "good competence in the subject". What is not made clear, however, is what a good grade is defined in relation to. The fact that there is no clear definition of what it will take for a student to reach this standard is a threat to the validity of assessment (Gordon Stobart, 2012), as it means the teachers will have to define that for themselves. Another key aspect is the collection and interpretation of evidence during the assessment, as this evidence is used to justify one's interpretation of the student's overall performance. As examinations happen in real-time, this evidence will be restricted to what the examiner notices in the moment, which can lead to different examiners noticing different things.

Nusche et al. (2011) comment that while the broad competence aims are meant to give teachers more autonomy when it comes to planning their teaching program, this freedom also causes challenges. Teachers can find it difficult to turn such broad and thus rather vague aims into something concrete because of the way the goals are phrased. This further increases the chance of divergent understandings, as it depends to a large degree on teachers' subjective judgment.

This, in turn, can lead to divergent results when it comes to grading. Sandvik (2013) points out that the assessment context is influenced by several factors, including what education the teachers had, how much experience they have had with ratings in general and examinations in particular, as well as their views of learning in general. This can subsequently affect how the evidence of students' performance during the examination is collected and subsequently interpreted to reach a common understanding of the performance. This is especially important during an examination, where two examiners can have different chains of interpretation which then culminate in an understanding of the student's knowledge.

If there is a common understanding of what is to be assessed, on the other hand, then that would help ensure the reliability of the assessment. As noted by Johannessen (2018), "the competence aims are to be general and possible to assess at the same time" (p. 22), which is difficult to combine as long as there is no clear definition of what constitutes adequate performance and how it compares to good or excellent performance. As Bøhn (2016) points out, the meaning of a descriptor like "very good" in the assessment criteria is unclear as it is not defined what it is compared to. This baseline can be found in the subject curriculum, though the number of different competence aims there means that the meaning of the score given will depend on the assessment situation.

The second problem, as noted by Bøhn (2016), is that one can gather different kinds of evidence to support the assessment the teacher makes about the student's level of achievement. When a teacher is to interpret fluency, for example, the teacher might decide on a level based on the students' speech rate and pauses by measuring the number of times the students hesitate. The teacher might also use general observations of the students' talking to see any particularly noticeable occurrences of hesitation, or they might base themselves more on an intuitive feel for how good or bad the students' fluence is. The quality of the evidence gathered by each of the examiners can vary.

One solution can be to divide the competence aims into smaller objectives, which can make it easier for the students to understand what they need to do to perform at a certain level. However, that approach can also lead to an impractical amount of details, which may cause the connection between these smaller goals and the overall competence aims to be blurred.. (Hartberg, Dobson & Gran, 2012) To avoid this, the examiners should ensure they have enough good quality evidence, rather than trying to focus on everything the student does and subsequently being overwhelmed.

2.6 Equity

Cumming (2013) notes that, in a validation study, the impact which the test has on teaching and learning could be relevant. If one student receives a different grade for roughly the same performance as another student, this can have an outsized impact given the role which grades play in our society. If this difference were caused by one examiner believing that the construct of content is more important than that of communicative competence, then the examiners' operationalization will have played a critical role in what happened.

2.7 Previous Research

In this literature review, I will place this thesis in the larger oral assessment research context, using articles pragmatically selected because of their relevance to the themes of the thesis. The literature was chosen based on the need for further elaboration on the following topics: (i) What has been found about assessment in an international context, and (ii) what has been discovered about assessment in a Norwegian educational context. This combination will then be used to provide further insight into variability in raters' scoring and assessment behavior.

Before reviewing relevant research, four limitations to this literature review must be explained and clarified. First, most of the included studies use a different system to relate to assessment than the curriculum-based achievement assessments, which is what is used in the Norwegian system. The focus on systems of speaking proficiency tests can affect some of the perspectives in the studies, though the findings are still valuable. Second, many different age groups were involved in the studies rather than lower secondary school, which means a larger variation in skills. There are also varying degrees of proficiency and experience both when it comes to the testers and the raters. Thirdly, most of the assessments use a common rating scale, which is something not present in Norway as the scales for oral examinations are locally developed.

Finally, none of the reviewed studies have looked at what takes place during the examination situation itself, though their focus on the teachers' thoughts on what they are to assess is a very useful supplement to this thesis' observations.

2.7.1 Assessment Research in an international context

To make it easier to review the previous research, it has been sorted into three categories based on themes considered to be relevant to the present thesis. These themes are repeated in the conclusion of this chapter.

2.7.1.1 Lack of a shared understanding of assessment criteria

Kim (2015)'s study found that there was some variability in the behavior of raters. They looked at three groups of language teacher students and current language teachers, sorted by their level of experience when it came to rating, to see how their backgrounds might affect the way they use a scoring rubric to rate oral performance. These three groups were novices for those without any experience, developing for those with two or three years of experience, and experienced for those with over five years of total experience.

It was found that each group had a different understanding of the rating scale given to them, where novices were often confused by it while developing raters misunderstood parts of it less frequently than novices, and the experienced raters usually understood it correctly. Kim (2019) suggests that this might be caused by a limited understanding of the language concepts on which the scale was based, as well as a general lack of experience in assessing speech making it harder to assign a level of achievement to the student.

Ang-Aw and Goh (2011) found that examiners could award different scores to the same performance, as well as giving the same score to different performances. They noted that this could be due to different interpretations of the abstract terms in the assessment criteria. They also found that there was a pattern regarding how strict or lenient an examiner would assess a student. A strict examiner would consistently give lower grades while a lenient examiner would give higher grades. It is also worth noting that the stricter examiner disliked non-standard English, while the more lenient examiner focused on how it was used communicatively.

2.7.1.2 Subjectivity regarding the relative importance of each criterion

Iwashita, Brown, Mcnamara, and O'Hagan (2008) found that four features have an impact on the assigned score as a whole. These features were vocabulary, fluency, grammatical accuracy, and pronunciation, where the former two were viewed as especially important. Finding a link between pronunciation and intelligibility, they noted that this can end up affecting the whole performance since a teacher cannot look at other criteria if they can not make out what is said by the student. They also found that being weaker in one of these features did not always affect the overall rating, thus emphasizing the role which a holistic assessment played in the final rating.

Meanwhile, Orr (2002) conducted a study that suggested that variation in rater perceptions might also affect the assessment. Looking into the thoughts of raters while assessing oral performance in English, Orr found that raters focused on different aspects of the assessment rubric given to them. Because of this, Orr (2002) found that there were variation in the scores given to the student's performance and that raters could perceive the same performance differently and yet give the same grade.

Ang-Aw and Goh (2011) found that examiners tended to emphasize content over the quality of the student's language. Even though Singaporean teachers had gone through rater training before the test, they still found that there was some variance in the assessment given. This variation was caused by different degrees of views on whether all criteria in each factor had to be included, what part of the student's performance they focused on, and the method by which they combined it all into a final grade. This shows the role subjectivity can play, as the complex nature of oral assessment means that raters have to balance a lot of different parts of a performance and find a way to use it to grade them.

They also noted that, while the teachers said that they believed all features to be important, some criteria received more attention than others. The ability to give personal responses received the most evaluative comments, followed by comments on how well the student had done holistically. They interpreted this to mean that the examiners' final decision might be influenced by the one or two criteria they viewed as most important.

2.7.1.3 Influence of non-criterion-based aspects of performance

One problem suggested by research is that teachers who are assessing based on the same scoring rubric can come to the same grade for different reasons. David Douglas (1994) went through the transcripts of six students' tests, where they'd all received similar grades on their oral performance. Analyzing these transcriptions by looking at various common aspects used to assess performance like content and vocabulary. He found there was little correlation between their performance and the test scores given to them by the examiners. Based on these findings, he suggested that this discrepancy might be caused by the teachers' assessment being influenced by criteria that were not found in the rater scales.

Sandlund and Sundqvist (2016) conducted a study looking into equity in L2 English oral assessment, where examiners first assessed their own students before collaborating with another teacher to co-assess them. The study focused on those students who received divergent grades from their teacher and the external examiner. One thing they noted was that students of slightly different proficiency levels might end up being compared against each other rather than just against the assessment criteria, which might cause differences in proficiency to stand out more. They also note that the abstract phrasing of assessment criteria will leave teachers with personal interpretations of them and that teachers might come to value some approaches higher than others. They also point out the possible consequences of the student's teacher being one of the people assessing them. The examiner knows the students, which might lower anxiety, and their opportunities to assess the student earlier could make it easier to overlook minor mistakes. However, the examiner might also have preconceived notions of the student's level of performance, which might cause them to overlook it if the student actually performs better.

Jenkins and Parra (2003) looked at the role which nonverbal features plays during oral tests with four Spanish-speaking and four Chinese-speaking teaching assistants. They found that two types of behavior influenced the examiners. The first of these types were non-verbal and included behavior such as eye contact or the lack thereof as well as the use of gestures, head nods, and other types of positive body language. The other type was paralinguistic features, which included vocal qualities not always tied to linguistic features, like the rhythm and speed of the students' speech as well as articulation. (Jenkins & Parra, 2003). Their findings were that if speakers employed non-verbal behavior which was considered fitting for the test situation, they generally received a higher rating than those who did not employ such strategies.

Such variations in rater behavior were also found in a study conducted by Ang-Aw and Goh (2011). One examiner only used comparisons once while others used them more often, and one rater adjusted the mark they gave based on such a comparison. Ang-Aw and Goh noted that this comparison risked turning the criterion-referenced test into more of a norm-referenced one, where the grades are given would be affected by the performance of those who had gone before them. The researchers believed that this was caused by ambiguity in the descriptors of the criterion, as comparisons were used to establish a baseline understanding of what level each descriptor referred to. They also found that if the performance did not match up with the descriptors well enough, raters might end up feeling that a particular grade is unsuitable for that performance.

2.7.2 Assessment Research in the Norwegian Context

Henrik Bøhn (2016) looked at what was assessed by teachers by looking at the performance aspects the teachers focused on during an oral English examination in upper secondary school, Bøhn found that teachers generally focused on two constructs during the assessment. These constructs were communication and content. After organizing the different aspects which were observed, Bøhn noted that these main constructs consisted of several subcategories. The most important of these were "linguistic competence", which was sorted under communication, and "application, analysis, and reflection", which were put under content. The sub-category of linguistic competence could be further subdivided into grammar, vocabulary, and phonology where the latter was found to be the most important in the grouping. This shows that each category contains several aspects, and it is important to be aware of how teachers understand all of these.

Bøhn (2016) also found that teachers tended to have a shared understanding of the constructs and the sub-categories which were to be assessed, but there was still some variation with regard to what the individual teachers viewed as the most important of the categories. An example of this is how there were different understandings of how to weigh the construct of "Content" as a whole compared to the other construct of "Communication". There was also disagreement on what was most important of the sub-categories, with one example being related to phonology as some examiners viewed nativeness as a key factor (Bøhn, 2016). Still, they were generally in agreement that intelligibility was to be the most important factor in that subcategory. Finally,

Bøhn (2016) found that teachers tended to focus on language during the examination of students with lower levels of proficiency, while the content was viewed as more important for students with a higher level of understanding.

An example of an irrelevant performance feature brought up by the teachers is effort. When students with a weaker level of proficiency in the language, who were at risk of failing, were to be assessed, the teachers occasionally give them credit for trying their best. Still, Bøhn (2016) says that this was not universal, and some teachers explicitly took the opposite stance by arguing that they could not give credit for effort since that was not part of the assessment criteria they were given.

To research what performance features teachers focused on during English oral examinations, Yildiz (2011) interviewed a total of 16 teachers in upper secondary school. Her data showed that teachers utilized a variety of different features to assess the totality of a student's performance. These features were categorized as language competence, communicative competence, subject competence, ability to reflect and discuss independently, and finally, the student's ability to speak freely and independently of any manuscript. Yildiz (2011) also found evidence that suggested that teachers pay attention to non-assessment-based criteria. There was also some variety in what criteria were used and how some were viewed as important than other criteria, which Yildiz (2011) noted could threaten the validity and reliability of the examination.

Aalandslid (2018) conducted a study to look into how students and teachers understood oral competence. This was done by picking select competence aims and presenting them to teachers and students, alongside questions to elicit responses. She found that students often struggled to understand the competence aims, and they viewed the assessment criteria are more important to focus on. She also found that all the teachers focused on avoiding breakdown in communication and that the criteria they viewed as most important were pronunciation, fluency, and intonation. Additionally, the teachers viewed Norwegian-accented English as a threat to intelligibility and thus a potential cause of breakdown in communications, despite viewing nativeness as unnecessary.

2.7.3 Summary

The research articles reviewed in this chapter show that there is evidence of variability when it comes to how raters assess and grade the students' oral performance. This variability occurs because of one of three factors:

1. A lack of a shared understanding of the criteria that are used during the assessment.
2. Subjectivity with regard to the importance of each criterion
3. The influence of irrelevant performance features

There are challenges related to how irrelevant performance features are being used during assessment, particularly how inter-candidate comparison is utilized to establish a shared understanding. There is also some variation in the teachers' understanding of the rating scale used, and the lack of a common rating scale for all schools impact this negatively. There are also problems associated with how some criteria are used more than others, and how some are not used at all. Raters can also weigh different aspects of each criterion differently. as well as viewing some criteria as more important within their overarching construct or viewing one construct are more important than another construct. These are the factors mentioned as possible reasons for the variability observed in the rating of assessment by teachers.

3.0 Methodology

In this chapter, I will present the methodology that I have deployed in order to answer my overarching research question: How do teachers operationalize the assessment criteria during an oral examination in English in lower secondary school? First, I present the ETOS project, which my thesis is a part of (3.1) before I describe the research design I have chosen (3.2). Then I present the sample and the sampling procedures I used in the selection of participants (3.3). Next, I address the data collection procedures and the data material (3.4), before I outline the data analysis (3.5). Finally, research credibility and ethics will be discussed (3.6).

3.1 ETOS project

I was invited to become a part of the research project ETOS – *Evaluation of Bilingual Training Opportunities in Schools* in the academic year of 2021-22 (Brevik & Doetjes, 2020). ETOS is a project that aims to investigate bilingual teaching in two lower secondary schools in Norway, to increase our knowledge of bilingual education, where instruction is given partly in Norwegian and partly in English. The project also looks at the impact bilingual teaching has on students' learning outcomes, motivation, and relevance across subjects, including English. Among a variety of data sources, they recorded oral mock examinations. The project was initiated in 2019 and continued through 2022. It was led by project leader Lisbeth M. Brevik and deputy project leader Gerard Doetjes, with the former being my co-supervisor. The project received approval from the Norwegian Centre for Research Data (NSD/Sikt), and all the participants gave written informed consent prior to data collection, and explicitly stated whether they consented to their data being used for master's theses. All students and teachers who provided the material I use in this thesis, consented to such use. The sample in the project consists of two schools in areas with different socioeconomic statuses.

The data material I used in this thesis consisted of recordings from mock examinations at one of the schools; in comprised audio recordings of five students in total, with the same two examiners throughout. The data was collected using one dictaphone, placed near the students and the examiners.

3.1.1 On the reuse of data

Using data gathered by others for your own research is a choice with both advantages as well as potential drawbacks. Anderson-Bakken and Dalland (2021) note that the reuse of data can result in more information being obtained as the material is looked at from different perspectives and in a more thorough manner than if it was just done once.

Additionally, there are timesaving and economic concerns that are of some importance to master students, as they do not have to spend time and money to go out and record, nor do they have to go through the process of finding and recruiting informants. Reusing recordings also means that there is less of a need to go out and record, which means that the disruptions involved in recording a class can be avoided in some cases. This decreases the pressure on schools from researchers looking to obtain data for their projects.

One potential drawback is that the data might not fit as well as data material recorded specifically for a given purpose. I have attempted to limit this by talking to the original researchers about the exact purpose and role of the recordings, which makes it easier to ensure that the recordings are fit for the intended purpose.

3.2 The research design

Johannesen defines research design as "How a study is organized and conducted to make it possible to answer the research question" (Johannesen, et al., 2005). There are a number of possible qualitative research designs to choose from. When deciding which one to use, the researcher must consider the purpose of the study.

I wanted to understand how examiners interpret the assessment criteria during oral examinations in lower secondary school. My study focused on a limited number of cases related to a phenomenon, which means that it is categorized as a case study (Johannesen, et al., 2005).

3.2.1 A qualitative or quantitative study?

It was clear from the beginning that I wanted to do a qualitative study, as I wanted to find out more about how teachers agreed on how to assess an examination, as well as what aspects of a student's performance or which criteria they focused on when grading said examination. The qualitative approach aims at describing and characterizing, providing extensive information

about the topic. This approach, being particularly suitable for the examination of phenomena we know little about, thus suited me well (Johannesen, et al., 2005; Bakke, 2010).

I was interested in observing what examiners actually do during examinations and comparing that to information about reported teachers' views on examination as presented in theory and prior publications to see how they compare. Because of that, I have chosen a qualitative research design, as that allows me to investigate how this is achieved in the classroom and the examination room of a specific class. A qualitative approach is well suited for the task of describing a phenomenon the way it is understood and evaluated by people (Ragin, 1994). This makes it useful for finding out how examiners operationalize the criteria.

Qualitative studies are well suited for investigations of a topic and to develop questions to figure out the "how" and "why" of something. While qualitative and quantitative studies cannot be completely separated from each other as they overlap in certain things, like how qualitative studies can include the counting of how many instances a specific action or code occurs in the material (Andersson-Bakken & Dalland, 2021), it is nonetheless clear that this study is mostly on the qualitative side of the spectrum.

I wanted to look at the actions of the participants themselves, to see how they operationalized assessment criteria in a realistic situation. Because of this, I found qualitative methods most suited to answer my research question. While there has been research into *what* teachers assess, there has been less focus on *how* they assess (Bøhn, 2016). This lack of research made it even more interesting for me to look into.

Table 3.1 gives a brief overview of my research design, including the overarching research question, the methods I have used, the data material and analysis, and analytical concepts.

Table 3.1. Overview of the research design for my MA thesis

Research Question	Research Design	Data Material	Data Analysis	Analytical Concepts
How do examiners operationalize assessment criteria	A qualitative study of examiners'	Audio recordings of five students'	Mixed, deductive and then	1: Assessment criteria-based construct

during an oral examination in lower secondary school?	discussions when assessing student performances at mock examinations	oral examinations	inductive approach	(Content and Language) 2: Non-criterion-based construct

3.3 Sampling

The ETOS project used purposeful sampling (Creswell, 2007) in the sense that the schools and classes were recruited on the basis that they offered bilingual programs at the lower secondary level. While the project collected data from grade 8, 9 and 10, only in grade 10 was data collected from oral mock examinations, with a total of two examiners (n=2) and five students (n=5) being involved in the examination situation. All participants had consented to participate in the research project, both before and after the recordings (Brevik & Doetjes, 2020)

Table 3.2 Overview of data material

Method	Data	Participants	Quantity
Qualitative	Audio recordings	Examiners (n=2) Students (n=5)	Recordings of five student examinations

3.4 Data collection and material

In this section, I will briefly explain the ETOS standards and procedures deployed to collect the data I have chosen to use in my study (Brevik & Doetjes, 2020). I will also include certain aspects of the collection process that I believe are of importance to the result of the study. I choose to include this information to give the reader some insight into the data collection process. This will give a broader overview of the data collection when discussing aspects such as internal and external validity later in this chapter (see Section 3.6.). Giving the reader insight into the whole process also contributes to the openness and transparency of the study, which Befring (2015) notes help increase its legitimacy and overlap with what Creswell & Miller (2000) call thick, rich descriptions.

3.4.1 Data collection procedure

Preparation for data collection started in the autumn of 2019 at the University of Oslo before entering the research sites. The data collection was conducted in January and February 2020.

According to Silverman (2011), the focus of qualitative research is authenticity. Observation, especially through a recorded medium, allows for this, which might be why it is one of the most popular methods to use for qualitative research. He further notes the importance of not choosing too many data sets to answer a research question, when wanting to describe and interpret different sides of a phenomenon. Thus, a smaller dataset like the one in this study makes it easier to go into the details of each recording as it is easier to get an overview and then go into the smaller, more granular details of each examination within the timeframe available for a master thesis.

3.4.2 Examination Audio Recordings

The research data gathered during the examination was audio recordings, with the recording being started and stopped by a research assistant who left before the examinations begun and came back to turn off the recorder and retrieve it once the examinations were done. As Brevik and Doetjes (2020) note in their report on the ETOS project, the use of one recorder was minimally invasive. Furthermore, and most importantly, each student and both examiners consented to the recording of the assessment situation (Brevik & Doetjes, 2020).

Strict procedures and standards established in the ETOS project with regard to the handling of data material were followed before, during, and after the recording (Brevik & Doetjes, 2020). All recordings were made using a physical dictaphone and encrypted hard drive. Each recording was transferred from the dictaphone to the encrypted hard drive, thus ensuring that best practices were followed to protect private information.

3.5 Data analyses

3.5.1 Transcribing the Recordings

My data consisted of transcriptions of the recorded mock examinations. By transcribing the interviews, I was able to get a clear and structured overview of the data, which made it more suitable for analysis (Kvale & Brinkmann, 2015). Before each transcription, I listened through

the recording once in its entirety. In Chapter 4, excerpts from the interviews will be used to illustrate the results.

3.5.2 Analyzing the Recordings

Analysis of qualitative data has to do with editing and analyzing (Johannesen, et al., 2005). In qualitative research, this means that you have to work through the data, compare the answers, and find similarities and differences, patterns, and other interesting elements (Johannesen, et al., 2005). I did this by separating the examination itself, the examiners' conversation afterward, and the feedback they gave to the student, before making an overview of the total amount of time different concepts were brought up. This includes non-criterion-based concepts. I also went through each examination to see which examiner brought up which criterion. This overview then made it easier to look for patterns and see which of my initial notes were validated by the data and which turned out to be a case of magnification of what were minor occurrences.

As my goal is to provide knowledge and understanding of the phenomenon being studied, I went with a combination of deductive and inductive techniques. In my first cycle of coding, I used the assessment criteria the teachers were given as the foundation for the codes, with one code for each assessment criterion. This was then supplemented by inductive coding of non-criterion-based concepts, as those did not have a basis in the assessment criteria. This analytical framework was then applied to all five examinations.

3.5.3 Key analytical concepts

The first thing I did with the data material after I had read through it was to divide it into two constructs. This was done to help sort the assessment criteria used by the examiners. One of these was language, the other was content. This division was done based on how content and language are defined in the English subject curriculum and the competence aims. I also looked at how they were defined by Bøhn (2016) for use in his doctoral thesis.

Below is a table that shows the analytical concepts used and an explanation that shows the difference between the two. It also shows the sub-groups' language and content, which each assessment criteria were sorted into.

Table 3.4:

Analytical Concept	Explanation	Example
Criterion-based constructs	Criteria are based on assessment criteria found in the document given to the examiners as well as the students in advance of the examination. Divided into two different constructs, language and content.	Vocabulary (Language) Understanding of the topic (Content)
Non-criterion-based construct	Criteria that are not based on assessment criteria yet were still used during the assessment to comment on the student's performance during the examination	Comparison with previous students, Creativity

Table 3.4 shows the two analytical constructs, the criteria based on assessment criteria and the criteria not based on assessment criteria.

Examples that seemed to be particularly noteworthy were marked with a yellow marker to make it easier to find them again for use as illustrative examples. It should be noted that the competence aims listed below are taken from the previous version of the subject curriculum (2006), after it was revised in 2013 (ref). This was the governing document in lower secondary school when the collection of the data material was conducted. Because the assessment criteria were not sorted into the categories of language and content in the document the examiners received, I sorted them myself.

3.5.4 The use of non-criterion-based features of the presentation

Whereas the sorting of the criterion-based assessment criteria was based on theory and on the competence aims provided by the directory of education, the non-criterion-based features were coded intuitively. In instances where the examiners discussed something about the students' performance that was not found in the assessment criteria, that utterance was marked with a red marker in the transcript. After all the transcripts had been analyzed and coded, I looked at similarities between what had been marked and sorted them into categories based on what features they referenced.

3.5.7 Restructuring of assessment criteria

The assessment criteria were translated into English by me and then given a code that was used to refer to them throughout coding and analysis. An example of such a code is *vocabulary*. In general, each individual code refers to one of the assessment criteria in the list used by the examiners. The assessment criteria were translated and not edited, with one notable exception. To avoid the potential confusion caused by having too many different parts of the students' performance together under one heading, one of the assessment criteria was split up to make it clearer what the examiners brought up during the examination. This split also helped avoid duplication since, otherwise, some criteria mention the same parts of the student's performance. The criterion that was split up was named *reading, evaluation, and comparison of texts*, which can be seen below.

Table 3.5

Low	Middle	High
Reads simple texts in different genres and about different subjects and retells these. Communicates a message through the use of simple words and expressions	Reads different texts and retells these. Presents timely themes with logical cohesion and talks about them freely based on some keywords.	Compares, talks about, evaluates, and presents several varied and timely topics in a purposeful way

Table 3.5 shows an assessment criterion that describes several different aspects of a student's performance. This means that it would be unsuitable as a code, as it would not be clear what it referred to.

The criterion that was split up had several sections. The section regarding the student's ability to "communicate a message through the use of simple words and expressions" was given its own code, as it did not fit in with other assessment criteria. This decision helped keep the different aspects of the student's performance separated into the appropriate construct, as reading and comparing texts belonged under content while communicating a message belonged under language. Additionally, "presents several varied and timely themes with logical cohesion" were moved to the *fluency and cohesion* criteria, while "talk about them in a free way based on some keywords" were moved to the *script* criteria.

Below is a table that shows the result of this process. Three different codes, based on other assessment criteria, received parts of it that fit with that code's theme. Below that is a table that shows the original assessment criteria used by the examiners during the examination. (Translation mine)

Table 3.6

Vocabulary	Communicates a message through the use of simple words and expressions
Fluency and cohesion	Presents timely themes with logical cohesion
Script	(…) and talks about them in a free way based on some keywords

Table 3.6 shows what other assessment criteria received the parts that were removed from assessment criteria named criteria *reading, evaluation, and comparison of texts.*

3.6 Research credibility

In this section, I will discuss the reliability and validity of my study, along with ethical considerations. According to Johnson and Christensen (2013, p. 278), validity refers to "the correctness or truthfulness of the inferences that are made from the results of the study", and that reliability is present "when the same results would be obtained if the study were conducted again (i.e. replicated)" (2013, p. 278) Further, Brevik (2015, p. 46) argues that the difference between the two concepts can be described as "the trustworthiness of the inferences drawn from the data (validity)" and "the accuracy and transparency needed to enable replication of the research (reliability)". For a study to have validity it must therefore have reliability; but a study can have reliability without having validity, which I discuss below.

3.6.1 Reliability, or repeatability

Johnson (2013) states that a study's reliability is concerned with how and if the results obtained are repeatable by other researchers. However, qualitative research is by its very nature impossible to repeat. As Brevik (2015) states, "Research, where people are involved, can never be fully replicated; for instance, the atmosphere in a classroom will never be identically recreated and identical utterances will not be uttered" (p. 46). Because of this limitation, it is

more typical for qualitative studies to adopt the social constructivist stance, which states that while research can never be replicated exactly the same way, it's still important to make it as transparent as possible. This is done by detailing the research process itself in as much detail as possible (Gleiss & Sæther, 2021).

Hallgren (2012) states that reliability can be divided into inter-reliability and intra-reliability. Intra-reliability is concerned with to what degree the study agrees with the results of other researchers. My study used a mix of inductive and deductive approaches, in part based on self-defined codes and in part on categories related to assessment criteria only relevant to a single school. While the codes have not been used by others before, steps have been taken to help limit the impact of this. The number of details in the analysis section of this thesis, and the inclusion of the codes as well as the assessment criteria themselves as an appendix, would have made it easier if other researchers were to try and replicate my findings.

Additionally, intra-reliability measures to what degree there is an agreement among multiple repetitions of one test (Bryman, 2016), which in this setting can be understood as the collection of data material. All the data gathered through the ETOS project were collected according to the project's standards, using the same equipment, which helps maintain a consistent level of quality (Brevik & Doetjes, 2020). The fact that the data has been accessed by several members of the ETOS team, allowed me to discuss my interpretations of the data with them, which allowed me to consider more viewpoints than I would have done by myself. The fact that they were recordings also made it possible to review the data multiple times, thus further improving reliability.

3.6.2 Validity

In this section, I give an account of what strategies I have employed to enhance the validity and trustworthiness of my study. Regarding validity, Johnson (2013) states for a study to be deemed valid, "it has to be plausible, credible, trustworthy, and therefore defensible" (p. 299). Brevik (2015) notes that validity does not refer to the data itself, but rather to the researchers' judgment and thoroughness through the process and finishing of a study. It also related to whether the conclusions and the inferences drawn from the data are trustworthy and defensible.

The transcriptions of the audio recordings of the examinations have been carried out by me and are available to the other members of the team to look at and review for themselves, thus adding to the transparency and the descriptive validity of the study, as readers can look at the sources themselves and compare that to the conclusions I have made based off of my interpretations of the data, and then decide to what degree they think that I have presented a trustworthy analysis of the information found therein. (Johnson, 2013). Additionally, doing the transcriptions myself gave me greater insight into the data material.

Creswell (2014) states that qualitative validity means that "the researcher checks for the accuracy of the findings by employing certain procedures" (p. 201). Therefore, the researcher cannot rely on the results alone, as they have to check if the findings from the study might, in fact, be wrong. To ensure that the findings I have are as correct as possible, I will go through two factors that might influence the validity of the study.

Reactivity relates to the influence a researcher might have on a setting or its people in a study (Maxwell, 2013). Firstly, an observer or a researcher being present in a setting might affect the participants, according to Kleven, Hjardemaal, and Tveit (2014). In turn, this could create an unnatural environment for the participants, preventing them from relaxing and acting as "themselves", affecting the results and inferences drawn from the data. While the data gathered in my study did not involve a researcher being present, a microphone was still present in the room. Wickström & Bendix (2000) argues that reactivity will also be present if electronic means of observation like recordings are used to collect the data material.

The reason for this is that the participants' awareness that what they are doing is being recorded and will be reviewed by someone can impact what they say or the way they phrase things (Wickström & Bendix, 2000). However, none of the students commented on the recording equipment at all. A reason might be that the students had been recorded in their regular classroom lessons for two weeks prior to the mock exam. They were also asked before the mock exam whether they consented to it being recoded. On the day of the exam, they were asked if they still wanted to be recorded, and they were reminded that they could withdraw their consent after the exam should they wish to do so (Brevik & Doetjes, 2020). In fact, the only times the examiners commented on the recording equipment were before their first discussion, and at the end when the research assistant came in to retrieve the recording equipment. Based on this, I would therefore argue that the audio recordings to a large degree

depict the natural environment of the examination conversation for the student as well as the examiner conversation where the students' performance was assessed. This fits well with the findings of Klette and Blikkstad-Balas (2017), who argues that this effect on the participants is not as significant as feared. Her research showed that people often forget that they are being filmed, and the same can presumably be said to be relevant for audio recordings as well.

The researcher's bias, which is to say the preconceived notions, and values a researcher holds related to the topic that is being researched, might influence the inferences I draw from my study. If so, that would end up affecting the results and the validity of my analysis (Maxwell, 2013). During the process of writing this thesis, I have attempted to minimize researcher bias as much as possible. This was done by trying to expect or at least be prepared for unexpected findings, as well as actively not searching for the results I assumed I would find, as recommended by Johnson (2013). This, combined with discussions of the findings with others, has helped me avoid being too influenced by my own biases, though they cannot be removed entirely.

3.6.3 Generalizability

As Bryman (2012) points out, it is impossible to generalize statistically from a small, non-randomized sample to a population. The sample size is too small, and not representative enough for this to be possible. Thus, the findings from the sample used as the data material for this thesis cannot be statistically generalized to the population of English lower secondary school examiners in Norway. However, other forms of generalization are possible.

Two types of generalizations that are relevant for my thesis are analytical, or theoretical, generalization (Gleiss & Sæther, 2021; Mitchell, 1983), and generalizable patterns (Larsson, 2009).

In this study the notion of theoretical generalization, and its arguments that "the cogency of the theoretical reasoning" (Mitchell, 1983, p. 207) are decisive for judging the interpretation of the results, is particularly relevant in relation to the content construct. During my analysis of the data, I found evidence that suggests that the teachers were largely utilizing the non-criterion-based constructs to agree on what level a student's performance was on when it came to whether it was correct to give them the highest grade or not, which is further empirical support for Ang-

Aw & Goh (2011)'s finding which suggested that teachers use of non-criterion-based criteria might be tied to vague assessment criteria making them search for more to base their inference on, thus suggesting that this is a relevant way of describing how teachers act when they assess a student's level of performance during an oral examination.

The idea of generalizable patterns, on the other hand, which can also be defined as "configurations, which can be recognized in the empirical world" (Larsson, 2009, p. 33), can be used for findings that are not so much a matter of theoretical representativeness, but rather of empirical resemblance, that is that they can be found in the real-life situations themselves. Thus, some findings like the general agreement among the examiners that both language and content are to be examined equally, the fact that some criteria were discussed often and others less so, and the role of non-criterion-based constructs in reaching an understanding of student performance may be seen as patterns. These patterns are transferable from the specific research situation, that of a mock oral examination intended to be as close to an authentic oral examination as possible, to real-life lower secondary exams and thus they are points which it might be useful to keep in mind by examiners preparing for their role (Yin, 2016).

3.6.4 Ethical considerations

Research ethics has played a major role in ensuring the privacy and well-being of the participants through the data collection, the processing of the data, and the writing of this thesis. During the data collection period, the ETOS team received firsthand experiences with how to protect the privacy of teachers and students who participated in the ETOS research project, in line with the GDPR requirements. GDPR, or General Data Protection Regulation, is a regulation in EU law on data protection and privacy, with a focus on the protection of personal data and the transfer of this. All participants were anonymized in this study, as each person was classified as either an examiner or a student based on their role. A number was then added for each person with that same role, like for Examiner 1 and Examiner 2.

Befring (2015) emphasizes the right to privacy for all participants, thus making it clear that a researcher cannot collect data at all costs. One student was not comfortable with having the grade she received on her exam on the recording, so any discussion of grades was deleted from that recording by the ETOS team, before I was granted access to the data. Prior to the data collection, all participating members of the ETOS project signed consent forms agreeing to

confidentiality regarding the project and its data, including myself. I did not play a role in the collection of data material itself, as the collection was done before I started my thesis and afterward the collection of relevant material was not possible due to the cancellation of all examinations following the Covid pandemic. However, I transcribed the recordings of the examinations. This was done through secure computers in the TLV lab (Teaching Learning Video lab) data lab which accessed the data material. The material was stored on an off-site server with access restrictions in place. These precautions, which included limiting access to only what people needed for their work, minimized the chances of accidental loss of data and ensured that the participant's right to privacy was respected.

3.6.5 Limitations

The audio recordings collected depicted a mock examination, rather than a mock examination. The reason for this was that the regular exams ended up being cancelled because of the Covid pandemic. While this means that they are not identical to an actual examination, they were nonetheless as close as it was possible to get during those times.

4.0 Findings

In this chapter, I will present my main findings based on an analysis of the data material. This data material consisted of the conversations the examiners had after each oral mock examination was completed about what grade they should give the student, and the conversation the examiners had with the student afterward when they informed said student about their grade.

First, the constructs and assessment criteria used during the assessment will be reviewed (4.1). Then the results themselves will be presented in three main parts: Results from the use of the language construct (4.2), results from the use of the content construct (4.3), and results from the use of non-criterion-based performance features (4.4). All sections will include data excerpts, which have been chosen to best illustrate the examiners' views and arguments during the examination. The examiners had the competence aims as well as the assessment criteria in front of them during the oral mock examination.

My first main finding is that the examiners apply a variety of assessment criteria to rate a student's performance, though which criteria are used and how many times they are brought up varies between examinations. My second main finding is that the examiners do not discuss the impact that language ideology can have on their operationalization of the assessment criteria. My third main finding is that the examiners generally agree on how to interpret the assessment criteria, but it is not always clear what level the student is at. My fourth main finding is that non-criterion-based performance features are brought up and discussed during the examiners' conversations.

4.1.1 Criteria used during the assessment

The data presented and analyzed in this section is an overview of the assessment criteria used during the examinations, as found in the five audio-recorded oral mock examinations. Explicit mentions of a criterion, like vocabulary, were counted as one single occurrence. If the examiners brought up a sub-criterion, like *subject-verb concord*, then that was counted as an occurrence of the assessment criteria to which the sub-criteria belong, which in this case would be *grammar*.

In his doctoral thesis, Bøhn (2016) found that teachers who taught vocational classes in upper secondary schools focused more on content than on language. To see if a similar phenomenon took place in the lower secondary school in my thesis, the assessment criteria were sorted into two central constructs, language, and content. This division would help reveal if the examiners focused more on one of these constructs. While these constructs were at first only intended to help sort the different assessment criteria, the analysis process showed that the examiners also referred to language and content directly during the examinations. Because of this, the number of times the constructs themselves were directly mentioned was counted as well. Additionally, a third construct was found during the analysis. This construct consisted of features of the student's performance that were not referred to in the assessment criteria yet were still used by the examiners. Such features are labelled non-criterion-based assessment features.

Table 4.1 presents the constructs used, the criteria related to each construct, as well as the number of time each criteria was discussed for each examination.

Table 4.1 – Overview of the three constructs and their associated criteria, as empirical data for each examination.

Constructs	**Criteria**	**OME 1**	**OME 2**	**OME 3**	**OME 4**	**OME 5**
Language		5	6	0	9	3
	Vocabulary	2	2	0	4	1
	Communication	1	1	0	2	0
	Fluency and cohesion	0	0	0	2	2
	Script	0	1	0	1	0
	Pronunciation	0	2	0	0	0
	Grammar	2	0	0	0	0
Content		6	5	1	8	4
	Understanding of the topic	1	3	1	7	3
	Reading, evaluation, and comparison (of texts)	4	2	0	1	0

	Reflection on questions	1	0	0	0	1
	Reading comprehension (Non-fiction)	0	0	0	0	0
	Sources	0	0	0	0	0
Non-criterion-based performance features		0	4	1	2	3
	Outstanding quality	0	0	0	1	0
	Effort	0	1	0	0	1
	Comparison	0	3	0	1	2
	Creativity	0	0	1	0	0
Total		11	15	3	19	10

Note: OME stands for mock examination.

As seen in Table 4.1, there is a certain degree of variation between the examinations regarding how many criteria were brought up by the examiners. The lowest number was observed during mock exam 3, where only a total of 3 criteria were brought up during the discussion. The highest number occurred during mock exam 4, where a total of 19 criteria were brought up by the examiners. Only one criterion was brought up for every single student, namely "understanding of topic and cause and effect".

It is worth noting that not all of the assessment criteria were brought up by the examiners during their discussions. However, the two criteria that were not used, “reading comprehension” and “sources” were not relevant, because the aspect of the student's knowledge they were meant to assess was not present in the exam. The fact that the criteria were not adapted to the task the students were given indicates that the assessment criteria the examiners were using were generalized.

Table 4.2 shows how many times one of the assessment criteria related to a construct was mentioned. This was done to see if one of the examiners prioritized either language or content more than their co-examiner did.

Table 4.2 – Overview of the examiners' use of the two main constructs during each examination. Non-criterion-based performance features were not counted.

	Content	**Language**
OME 1	5	9
Examiner 1	3	5
Examiner 2	2	4
OME 2	5	4
Examiner 1	2	1
Examiner 2	3	3
OME 3	2	2
Examiner 1	0	1
Examiner 2	2	1
OME 4	10	13
Examiner 1	6	8
Examiner 2	4	5
OME 5	7	5
Examiner 1	1	3
Examiner 2	6	2
Total	58	66

The data material shows that their arguments during the examiner conversation were not biased towards either content or language (58 versus 66 times). Below, I will go through the criterion under each construct to examine how they are used to come to an understanding of the student's level of achievement. Each quote was originally spoken in English unless otherwise noted. Throughout the examinations, the examiners focused and commented on different elements of the student's language. These criteria were sorted by frequency, with the most referred to criteria being shown first.

4.2 Language

4.2.1 Vocabulary

Vocabulary was the most referenced assessment criterion related to language. Both examiners referred to the students' vocabulary specifically, with both of them pointing to specific words which the students had used, which were viewed as representative examples of the student's overall vocabulary.

Low	Middle	High
Knows some English words and expressions and uses them.	Knows many English words and expressions and uses these to describe various subjects	Has a large vocabulary that is advanced, varied, and descriptive and can use in a purposeful way in different subjects with varied content

Figure 4.1. Level description for the *vocabulary* criterion.

Excerpt 1 shows an example of the examiners describing a student's good vocabulary, which included pointing to specific word choices viewed as indicative of this.

EXCERPT 1:

Examiner 1: Yeah she [student] does have some of her vocabulary [which] is quite, is quite good

Examiner 2: Mhm

Examiner 1: She's saying like 'betrayed', 'exposes', 'trauma'

(OME4, examiner conversation)

Excerpt 2 shows that examiner 2 referred to the student's level of achievement as 'excellent', and pointed out that the student displayed high-level vocabulary. Again, specific references to advanced words were made by the examiner.

EXCERPT 2:

Examiner 2: Excellent, so much good vocabulary

Examiner 1: Yes

Examiner 2: High-level vocabulary [is] essentially currently the best explanation for this
(OME1, examiner conversation)

Meanwhile, excerpt 3 shows an example where the students' vocabulary was found to be at a lower level of achievement. The examiners point out that the student's vocabulary lacks the expected range. They also gave the student a grade that they viewed as fitting for that level.

EXCERPT 3:
Examiner 1: No, I find that… uh… eh… he, her vocabulary: She doesn't have a wide…
Examiner 2: No
Examiner 1: …a wide vocabulary
Examiner 2: No, but no, I agree
Examiner 1: Yeah and she she's not… she's usi… she's not using a wide vocabulary
Examiner 1: Ahm so that that to me that was [stating a grade] vocabulary
(OME 5, examiner conversation)

Excerpt 4 shows the examiners commenting on an instance where the student asked them for the right English word after the student had said the word in Norwegian. The examiner seemed to view that as evidence of a lower level of achievement.

EXCERPT 4:
Examiner 2: And she's making the "cardinal sin" of asking us
Examiner 1: Yeah
Examiner 2: For words
Examiner 1: Yeah
(OME 2, examiner conversation)

It can be inferred from the examiners' conversation that not knowing the right English term will count negatively during the assessment, even if the student knows the correct term in another language. Thus, leaning on your L1 for support was not viewed as acceptable in this case.

4.2.2 Communication

Communication was the second most frequently used criterion related to language. Being able to communicate clearly at an even and understandable pace is important, as examiners cannot assess that which they did not hear. Likewise, an explanation that does not communicate what the student knows to the examiner, will negatively impact their grades.

Low	Middle	High
Puts together words so it makes sense and form understandable sentences	Expresses themselves with some precision.	Expresses themselves precisely

Figure 4.2. Level description for the *communication* criterion.

Excerpt 5 shows the examiners remarking that the student spoke quickly, and that this negatively impacted their ability to communicate their knowledge. This meant that the examiners experienced that they got a more limited view of the students' knowledge.

EXCERPT 5:

Examiner 1: And I think that rushing would, would [have] made her kind of, maybe, eh communicate [not] so clearly…

Examiner 2: Mm

Examiner 1: …some of her ideas

(OME2, examiner conversation)

Excerpt 6 shows the examiners when they are in the process of pointing out mistakes the student made. They also note that these mistakes were found to not be sufficiently disruptive, as they did not have a negative impact on her communication. This can be understood to mean that grammatical mistakes do not necessarily influence the student's ability to communicate, as long as the mistakes do not change the meaning of what they are trying to communicate.

EXCERPT 6:

Examiner 2: And using, you know, using lot of high level language; there are some mistakes that she makes, ehm, especially with preposition and word choices

Examiner 1: Yeah

Examiner 2: ehm, throughout, but nothing that impacts her communication
Examiner 1: No
Examiner 2: Eh she communicates well
(OME4, examiner conversation)

4.2.3 Fluency and cohesion

Fluency, the third most referenced assessment criterion related to language, is noteworthy because it is difficult to operationalize. Measuring fluency as how many stops or gaps there are in a student's speech is difficult and impractical, which means an examiner's understanding of it will relate to a more holistic assessment.

Low	Middle	High
Puts together words so it makes sense and forms understandable sentences	Expresses themselves with some fluency and cohesion. Presents timely themes with logical cohesion.	Utilizes a language with good fluency and cohesion

Figure 3. Level description for the *fluency and cohesion* criterion.

Excerpt 7 shows the examiners bringing up fluency to point out that the student is fluent. This is then further defined as her being very fluent further down in the quote. While her relative level of fluency is pointed out, however, it was not made clear what marks it out as being at a high level.

EXCERPT 7:
Examiner 1: She's fluent, she has good language, but maybe some of her content and reflection and understanding weren't as strong
Examiner 2: Yeah that's how I feel too
Examiner 1: Yeah
Examiner 2: So I...
Examiner 1: It's a hard one because…
Examiner 2: It is
Examiner 1: …because she's so, she's very fluent
Examiner 2: Mhm

(OME4, examiner conversation)

Excerpt 8 shows the examiners stating that the student had fluency, though it was not made clear whether they considered the fluency to be at a low, medium, or high level of achievement.

EXCERPT 8:

Examiner 1: Ehm, so that that to me that was [mentioning a specific grade] vocabulary but she has 'flyt' (Norwegian for 'fluency')
Examiner 2: Yeah there's fluency
Examiner 1: Yeah
(OME5, examiner conversation)

However, during the conversation they had with the student afterward, they bring up part of what marks a student as having good fluency. Excerpt 9 shows that good fluency was noted as being a key part of being proficient in the language and was characterized by the ease with which the student could speak with the examiners.

EXCERPT 9:

Examiner 2: Hi so we're going to give you a [grade]on, ah, your performance here today, ahm, language wise you are strong, you show good proficiency, good fluency with the English language, ahm, you you speak easy
(OME 5, student feedback)

The quality of the sentences themselves was brought up in another feedback situation, in excerpt 10, where the examiner viewed it as a sign of the students' high level of achievement.

EXCERPT 10:

Examiner 2: And complex sentences throughout so definitely "høy måloppnåelse" (high level of achievement) there
(OME4, examiner conversation)

4.2.4 Manuscript

The script criterion is interesting in that the goal is for the student to use their manuscript as little as possible. Yet, despite this, the students are allowed to bring a manuscript, presumably because it can help them remember parts of their presentation that they would not have otherwise remembered.

Low	Middle	High
Presents a topic by reading from a script	Uses a script in a free way, based on some keywords.	Speaks freely

Figure 4. Level description for the manu*script* criterion.

As a part of the examiners' conversation about OME2, the manuscript is brought up when they discuss what the student could do better next time they had an examination. Given that they recommend that a student bring notes next time, it can be assumed that the student was not able to speak freely about the topic. This could be because there were topics they could have spoken more about had they been better prepared when it came to structuring what they were saying.

EXCERPT 11:

Examiner 2: To that section, she also… and we can give her this tip: she needs to maybe make a list key vocabulary…
Examiner 1: Yes
Examiner 2: … [that] she wants to talk about
(OME2, examiner conversation)

Excerpt 12 shows an instance where the student's ability to speak about the topic without the use of a script to support her is brought up as a good aspect of her performance. In general, the ability to speak freely without a script is viewed as important to the examiners, as long as it doesn't negatively affect other parts of the student's performance, like their vocabulary.

EXCERPT 12:

Examiner 1: She's not maybe at the excellence level
Examiner 2: Mhm

Examiner 1: But meanwhile, her, like, she's speaking freely, she's …
Examiner 2: Mhm
(OME4, examiner conversation)

4.2.5 Pronunciation

Pronunciation is among the least referenced criterion by the examiners, but still important. In particular, an examiner's view on pronunciation is likely to be impacted by their view on language ideology. An examiner who views Norwegian-accented English as equally good as a General American pronunciation will assess it differently than one who prefers a Received Pronunciation accent.

Low	Middle	High
Switches a little between English and Norwegian pronunciation	Has mostly correct pronunciation	Has a very good pronunciation

Figure 5. Level description for the *pronunciation* criterion.

The assessment criteria for pronunciation view occasional switches between English and Norwegian pronunciation as the definition of a low level of achievement. In excerpt 13, a student's use of their L1 is brought up by the examiners. They pointed out that she pronounced words in a manner more reminiscent of how the words were pronounced in Norwegian than in English. This was viewed as a weakness. Additionally, it was noted that the words which the student pronounced incorrectly included key terms like NATO, which the student was supposed to know. The examiners viewed it as a more serious mistake if the word the student pronounced incorrectly was key to the message that the student tried to communicate.

EXCERPT 13:
Examiner 2: And practice pronunciation
Examiner 1: Yeah yeah
Examiner 2: She had, she has, ehh, well right now it […]
Examiner 1: Yeah

Examiner 2: Elicit, […] NATO, July, ancient, satellite
//Pronounced with a Norwegian-influenced English pronunciation
Examiner 1: Yeah
Examiner 2: Key words she was using
Examiner 1: Mm
Examiner 2: That she's pronouncing in Norwegian
(OME2, examiner conversation)

4.2.6 Grammar

Grammar is the least referenced assessment criterion, and was only brought up to point out that the student made subject-verb concord mistakes, a mistake considered noteworthy by the examiners.

Low	Middle	High
Has some familiarity with grammar	Has good grammar	Has correct grammar

Figure 6. Level description for the *grammar* criterion.

Excerpt 14 shows an instance where the examiners point to specific aspects of the student's grammar that they viewed as particularly noteworthy, namely subject-verb concord. The examiner repeats specific sentences that the student had said incorrectly, which revealed that the student did not seem to have internalized the rule of subject-verb concord. The examiner then points out that it is strange that the student had not mastered the concord rules, considering how strong their language was in general.

EXCERPT 14:
Examiner 2: He had a few subject-verb agreement problems
Examiner 1: Okay
Examiner 2: That I notices: jobs that pays more, treatment that ah have been given
Examiner 1: Yeah
Examiner 2: […] it was odd, because his language is so strong
(OME1, examiner conversation)

The topic of subject-verb concord was brought up later, as shown in excerpt 15, when the examiners were trying to agree on what grade they should give the students' performance. The student's struggles with concord were brought up as an argument as to why the student should not be given the highest grade possible, which can be interpreted as more proof that the examiner views this mistake as a serious one. What is viewed as serious grammatical mistakes associated with a lower level of understanding can thus affect the student negatively.

EXCERPT 15:

Examiner 1: I, if, if, that's, if that's, ahh, a 6 presentation I'd be really surprised how we could… he did have some of those concord mistakes with the subject-verb agreement…
Examiner 2: Mm
Examiner 1: …As you mentioned, but I, I didn't catch them as soon as you did, would, but you're the [external examiner], what do you think
(OME1, examiner conversation)

4.3 Content

While the content section of the assessment criteria only had three assessment criteria, which is half as many as there are in the language section, the content assessment criteria are the ones brought up most often by the examiners. Thus, it also plays an important role in the overall assessment of the student's total level of achievement.

4.3.1 Understanding of the Topic

Understanding the topic was the most referenced criterion, used when the examiners were discussing how much the student had understood of the content they were expected to know.

Low	Middle	High
Talks about the topic, with some signs of repetition from memory	Have obtained an understanding of the topic, and mentions causes and consequences	Have obtained a good understanding of the topic, and reflected on the causes and consequences

Figure 7. Level description for the *understanding of topic* criterion.

Excerpt 16 shows the examiners commenting on the student's knowledge of the topic she had explained to them. While she could talk generally about the cold war and what happened during that conflict, they point out that she did not seem to understand what capitalism and communism are. This can be interpreted as the examiners pointing out that it is a serious weakness when the student lacks the requisite knowledge of key terms they're supposed to explain.

EXCERPT 16:

Examiner 2: She also did a really good job talking about the Cold War just here at the end
Examiner 1: Yeah she did
Examiner 2: As well when she pulled out the timeline, but when she mentions the ideologies and then…
Examiner 1: Yeah
Examiner 2: …can't comment on…
Examiner 1: No
Examiner 2: …ah the profound differences between…
Examiner 1: Yeah
Examiner 2: …communism and capitalism.
Examiner 1: Yeah
Examiner 2: That that was a weakness
Examiner 1: Yeah, definitely
(OME2, examiner conversation)

Excerpt 17 shows an instance where the examiners focused on two parts of the student's knowledge. The first was the topic she had been told in advance that she was supposed to talk about, and the second part was her general knowledge of history. Thus, the examiners looked at both topic-specific and more general, yet subject-relevant knowledge to find out how much the student knew.

EXCERPT 17:

Examiner 2: Ahm but I, I felt like her general knowledge…
Examiner 1: Hm
Examiner 2: …and her understanding of the topics that she was expected to talk about…
Examiner 1: Yeah
Examiner 2: …ah was weaker than what we've seen
(OME5, examiner conversation)

Moving on to Excerpt 18, the examiners noted that the student did not seem to remember where the main character lived in the book. This was used as an example of the student not having sufficient knowledge of what happens in the book. This shows that the examiners can focus on both larger overall aspects like a whole ideology, as well as more granular aspects like whether the student remembers a specific place name.

EXCERPT 18:
Examiner 1: No, and ahm, and just, I think she confused Garden Heights
Examiner 2: Yeah
Examiner 1: Until we directed her…
Examiner 2: Mm
Examiner 1: …in the right way. She didn't know Garden Heights was in the…
Examiner 2: She did not know enough about the book
(OME 2, examiner conversation)

4.3.2 Reading, evaluation, and comparison of texts

Reading, evaluation, and comparison of texts were among the most referenced assessment criteria. It was brought up when the examiners were discussing whether the student had in fact read the book they were supposed to, or to point out how well they used the book while bringing it up.

Low	Middle	High
Reads simple texts in different genres and about different subjects and retells these.	Reads different texts and retells these.	Compares, talks about, evaluates, and presents several varied and timely topics in a

		purposeful way

Figure 8. Level description for the *reading, evaluation, and comparison of texts* criterion.

Excerpt 19 shows an instance where the examiners pointed out that the student had not read enough of the book *The hate u give*, that they were supposed to be prepared to talk about. The fact that the student brought up events that only took place in the movie but not in the book itself was brought up to support this statement. The examiners viewed how prepared the student was as an aspect of the performance, including whether they had done the work they were supposed to or not.

EXCERPT 19:
Examiner 2: Ahhm, and "The hate u give"
Examiner 1: Yeah
Examiner 2: She, she was not
Examiner 1: I felt she, I think, I don't know how much she read the book
(OME2, examiner conversation)

In excerpt 20, we see an example where the examiner focused on how the student brought up the book as well as how he used it effectively when he answered questions. This can be interpreted to mean that the examiner focused on how relevant what was mentioned in the book was to the topic at hand, as well as how the student utilized the book in their argumentation.

EXCERPT 20:
Examiner 1: I thought [in] the first part, he was answering lots of questions, seeming reflective and open and kind of trying to look at it from different areas, and also bringing in, ah, the book "Absolutely true diary of a part-time Indian", which was he brought […] in quite effectively […].
(OME1, examiner conversation)

4.3.3 Reflection on questions

Reflection on questions was brought up one time, when the examiners pointed out that the student was able to answer the questions they were asked.

Low	Middle	High
Can answer questions in a limited way	Can answer questions and give some reasoning	Can reflect and give the reasoning for the answers

Figure 9. Level description for the *reflection on questions* criterion.

Excerpt 20 also shows an instance where the examiner brought up the student's ability to answer questions. He noted that the student could answer them and that he did so in a way that was both reflective and open about his thoughts on the topic. Thus, the examiners showed that they looked at the students' reasoning around the topic, and not just whether they could repeat a memorized response.

Excerpt 21 shows the examiner noting that the student was able to answer questions, without commenting on the quality of the students' responses or reflections.

> **EXCERPT 21:**
> *Examiner 1: She's able to answer questions and use really good 'flyt' with her language and such*
> *(OME 5, examiner conversation)*

4.4 The role of non-criterion-based performance features

As noted by Brown (2004), the reliability of an assessment situation was dependent on whether the examiners applied the same standards to all students, and if they succeed in minimizing the potential influence of human errors, inherent biases, or subjectivity. Despite this, it has been shown in prior research that examiners have focused on various non-criterion-based features of the student's performance and that they have viewed said features as important during assessment situations (Bøhn, 2016). This could weaken the reliability of an assessment situation, as such factors would differ from examination to examination. Below is an overview of the non-criterion-based performance features observed during the examination, as well as the context in which they were used.

4.4.1 Creativity

The performance feature of *creativity* is defined by the examiner pointing out the students' special approach to the topic, and how this was different from what other students had done because of its original nature.

Excerpt 22 shows an instance where the examiner brought up the uniqueness of the student's ideas, even noting that the student was willing to take risks by bringing up personal ideas. The examiners informed the student that this was a positive aspect of their performance, which showed that the examiner viewed unconventional approaches to a topic as something that can strengthen the overall level of a student's performance.

> **EXCERPT 22:**
> *Examiner1: We, we could have talked about the cold war. Some of the ideas you brought in there and even the… you know, you bring the, yeah, the ideas you brought in were just original, original thought provoking, ahhm, creativity [and] risk-taking was just something that you're really, really good at*
> *(OME3, feedback to the student)*

4.4.2 Outstanding quality

The performance feature of outstanding quality is defined as the examiner pointing out that the students' performance was outstanding, or that their presentation was not at the highest possible level.

Excerpt 23 shows an instance where the examiner noted that the student's level of achievement did not deserve the highest grade, because it lacked some kind of outstanding quality. More specifically, the examiner noted that the analysis was not found to be at an outstanding level. It is worth noting that outstanding quality was not mentioned in the assessment criteria.

> **EXCERPT 23:**
> *Examiner 2: Ahm, but I don't feel like this is 6 level*
> *Examiner 1: No*
> *Examiner 2: It didn't have the kind of outstanding quality that I'm looking for*
> *Examiner 1: Mm*

Examiner 2: In the 6-level student
Examiner 1: In terms of some of its reflection or what?
Examiner 2: Analysis maybe, analysis of this historical detail…
Examiner 1: Yeah
Examiner 2: …which those who've been at the at the highest level have been able to do (OME4, examiner conversation)

4.4.3 Comparison

The performance features of *comparison* is defined as the examiner comparing one student's performance to that of another student, which was done at the end of Excerpt 24. This could be done to point out that the current student did better than earlier ones, or to point out that they were weaker than earlier students had been. This could help them reach a shared understanding because they had both seen the previous performances together. However, because we have a criteria-based assessment system, this should not occur.

Excerpt 23 shows an instance where the examiners noted that the students' analysis was found to not be at the same level as that of other students found to be at the highest level. The examiners' definition of what performance deserved the highest level of achievement seemed to be influenced by what previous students had been able to do in terms of unique analysis. Thus, what the student needed to be able to do to be given the highest grade seemed to be defined in part by what other students had done earlier.

Excerpt 17 is also an example where the student's level of knowledge is stated to be lower than that of earlier students. This shows that the examiner had the performance of earlier students in mind, and consequently that a student's level of understanding could be assessed in relation to whether it is better or worse than what came earlier.

In sum, these non-criterion-based performance features seemed to be brought up to fulfil three different functions. These functions were:

1. To note if the students' presentation had some particularly outstanding aspect which the examiner expected of students at the highest level of achievement,
2. To describe something viewed as important to the examiners which are not referred to in the assessment criteria.

3. To help establish an understanding of what level the student's performance should be sorted under, through comparison with how other students performed during the test.

4.5 Summary of the results from the different constructs

In summary, throughout the examinations, there was a general agreement regarding what level of achievement each student was on. Firstly, the examiners generally agreed both on what level the student was at and what grades were appropriate for said performance. There were no major disagreements or discussions. The number of times each criterion was discussed differed from examination to examination, depending on what the examiners viewed as most relevant to discuss.

The examiners' operationalization of the assessment criteria includes specific references to what the students said (as took place during their discussions of the student's vocabulary in excerpt 1) and discussions of the relative level of the students' analysis (as happened during their discussions of the students' analysis in excerpt 16). *Vocabulary* and *communication* were the most used assessment criteria under the language construct.

The examiners focused on two aspects of the student's vocabulary, both how broad the students' vocabulary was as well as whether they knew certain advanced words. When it came to communication, the examiners noted the importance of speaking at a steady pace without rushing, as that might affect the clarity of the ideas they're trying to get across. They also noted that grammatical aspects like prepositions and word choices were not in themselves sufficient to impact communication.

The most used criteria under the content construct were the students' *understanding of the topic* as well as *reading, evaluation, and comparison of texts*. With regards to the student's *understanding of the topic*, the examiners focused on the student's specific knowledge about the topic they'd been told to prepare themselves to talk about, as well as their general knowledge about other subjects related to the topic. They brought up higher-level topics, like what the student knew about a large ideology and lower-level topics like whether the student remembered the correct place-name where the characters in a book lived. When it came to *reading, evaluation, and comparison of texts*, the examiners looked at whether or not the students had read the text or not, how relevant the text they brought up is to the question they were asked, as well as how they utilized it in their argumentation.

Secondly, the examiners did not seem to discuss the impact that language ideology could have on their interpretations of the assessment criteria. With regards to pronunciation, at least one of the examiners did not view it as acceptable for students to utilize the vocabulary they had in their native tongue (L1) to help when they did not know how to pronounce the word in English. During their discussion of grammar, it was noted that subject-verb concord was a low-level mistake to make, though some models of grammar view this as unimportant. Despite these differing views on what was viewed as a sign of low achievement, this was not brought up.

Thirdly, the examiners rarely used the language of the assessment criteria during the examination. Instead, they mostly used value-loaded language like ‘excellent’ or ‘good’ to describe the students' performance while they agreed on a shared understanding. Language use included gradually more references to assessment criteria during the examination, as the examiners eventually started to use specific grades, like 5 or 6, to help define what they assessed the student's level of performance to be at.

Finally, there was some variance from student to student. Non-criterion-based performance features were brought up by the examiners while discussing some students but not others. These performance features generally fulfilled certain functions and enabled the examiners to point out the strong sides of the performance or to note what they viewed as a weakness. This variation might affect the validity and reliability of the assessment situation, which will be discussed in the next chapter.

5.0 Discussion

In this chapter, the results from the observation of the mock examinations will be discussed in light of this thesis's theoretical background as well as its relation to previous research in the field. First, the construct of *content* will be discussed in light of the examiners' comments on the assessment criteria sorted under it. The reason for this is that the focus given to content indicates that is viewed as important despite having fewer criteria than the language construct (Section 5.1). Second, what teachers focus on during assessment will be discussed (Section 5.2). This includes more granular aspects of the performance, like verb-subject concord, as well as more holistic aspects like fluency. The role of reliability as well as the balance between having an overview and having a focus on details will also be discussed. Lastly, the contrast between teacher autonomy and standardization will be discussed with regard to reliability (Section 5.3). The examiners' reference to non-criterion-based aspects of the student's performance during the examination will also be mentioned. Throughout the discussion, the need for the examiners to have a shared understanding of the assessment criteria that they are using to assess will be emphasized, as this has important implications for the assessment situation.

5.1 The Role of Content

The analysis revealed that *content* was the construct that the examiners focused on. It was the construct with half the number of criteria that language had, while still receiving roughly the same number of comments in total. This combination suggests that content plays a larger role in the examiners' focus than language does. This is noteworthy since, as Bøhn (2016) points out, there is very limited theoretical support when it comes to the analysis of content during oral assessments. The disproportionate amount of comments per criteria implies that the examiners view the content criteria as the ones in need of discussion and clarification.

It is not immediately clear if their focus on the construct was because it was viewed as more important, or because its criteria were viewed as harder to define and thus in need of more discussion. The first interpretation fits with Yildiz (2011)'s findings that teachers used different criteria, and also weighed them differently. There was also variation in whether the examiners focused on specific details related to each assessment criteria or if they focused on obtaining an overview of the student's competence in that area, which will be discussed in Section 5.2.

The three criteria related to the content construct were *understanding of the topic*, *reading, evaluation, and comparison of texts,* and *reflection on questions*. There were considerable differences when it came to the number of comments related to each criterion. *Understanding of the topic* was by far the most discussed criterion as it was brought up a total of 15 times, with seven of those times being during a single examination. *Reading, evaluation, and comparison* were brought up a total of 7 times, with 4 of those times being during one particular examination, though not the same as the previous criterion. *Reflection on questions* was brought up 2 times, during 2 different examinations.

5.1.1 Subject curriculum and assessment criteria

Related to the assessment criteria mentioned above, it is relevant to look at the English subject curriculum (UDIR, 2013b). What students are taught can vary widely from school to school and classroom to classroom, because of the lack of a standardized list of content that all teachers are mandated to teach their students. While this relative autonomy enables teachers to teach what they view as most relevant and most suited to the general knowledge and interests of the class in question, it also means that teachers are likely to have different conceptions of what constitutes a good understanding of the topic (Eurydice, 2008).

This problem is potentially complicated by the presence of an external examiner from another school. This examiner comes from a different local context and may have to adapt. However, while teachers are given broad latitude when it comes to deciding what to teach, they are expected to teach topics that are explicitly mentioned in the centrally written competence aims. It can be assumed that the examiners' awareness of how different teachers can interpret content is in part why content is discussed so many times. It is not necessarily more important, as it might be because they are seen as vaguer by the examiners and thus in need of more discussion.

5.1.2 Understanding of the topic

Adjectives play a key role in the assessment criteria, as it is what the examiners must operationalize. One or more adjectives, like "good" or "nuanced", describe the students' performance to help differentiate between different levels. (Bøhn, 2016). An example of the adjectives used to separate one level of achievement from another can be seen in the difference between a low, middle, and high level of achievement in the criteria of *understanding of the*

topic. This was the most discussed assessment criterion, which indicates that it was viewed as particularly difficult to operationalize.

It is worthwhile to look at how the assessment criteria are written to understand this further. The description given for a low level of achievement is "Talks about the topic, with some sign of repetition from memory". The description given for middle and high is similar to each other, with a medium level of achievement being defined as "Has an understanding of the topic and mentions causes and consequences" while a high level of achievement is "Has a good understanding of the topic and reflects on causes and consequences. The only difference between a medium and a high level of an achievement is how the examiners interpret one adjective, "*good*", as well as how they understand the verb "*reflect*". Thus, how they interpret these words will affect how they assess the student's performance.

These terms are not easy to interpret, which leaves more up to the examiners' subjective interpretation. "*Good*" is a value-loaded term that is open for interpretation, since what one examiner view as good might be viewed as merely average by another examiner viewing the same presentation. It is not made clear what something is good in relation to, and this lack of a clear definition is a threat to the validity of an assessment (Gordon Stobart, 2012). The same can be said for the verb *reflection*, as the examiners would presumably want to define the quality of the student's reflection rather than merely stating that the student did reflect during the examination. Nusche et al. (2011) comment that these broad definitions are meant to give teachers autonomy, while also noting that this same freedom makes it harder to turn them into something concrete. This further increases the chances of different interpretations by the examiners, because it relies on their judgment to a significant degree.

The grading process can impact how examiners assess their students. It is worth noting that examiners are not merely expected to state what level of achievement the student is at. They are expected to give the student a grade ranging from 1 to 6, whereas the assessment criteria only have three groupings defined as either low, medium, or high. Thus, even if the examiners decide that a student's presentation fulfills the criteria for a high level of achievement, they still have to decide whether a 5 or a 6 is most appropriate for the student in question.

At the end of the examinations, the examiners started to assign a specific grade to some aspect of the student's performance. They started to pick a specific number, rather than using the

categories of the assessment criteria. This could be because the examiners wanted to be particularly clear about that particular student's level of achievement in those particular criteria, but it is also possible that it was a response to the vagueness inherent in the achievement criteria. Whether or not this was a new approach that would have been used on later candidates cannot be ascertained because this was done for the last students.

5.1.3 Reading, evaluation, and comparison of texts

While discussing *reading, evaluation, and comparison of texts*, examiners focused on whether the student had read the book, and how effectively they utilized it. When the student's knowledge seems to come more from the movie than the book, the examiners point this out. This lack of knowledge indicates that the student did not read enough of the book. This too is a question of degree, as found above, but it is a clearer case than the previous ones as not having read the book is a clear weakness when they are expected to interpret it.

It is important to look at how the students used the text they were given as well. The examiners bring up how the student not only brought up the book but also that they used it effectively while arguing for their interpretation. As Brown (1996) noted, tests which are supposed to measure multiple factors including speaking as well as reading comprehension are more challenging to assess. It is worth pointing out that the assessment criteria go from saying that students should be able to "read simple texts in different genres and about different topics, and retell these" to expecting students to be able to "compare, talk about, assess and present several different topics in a fitting manner". However, rather than focusing on whether the student fulfills each of these criteria, the examiners seem to focus more on having an overview of the student's use of the texts. This will be further discussed in 5.2.

5.1.4 Reflection on questions

When looking at the second criterion, *reflection on questions*, it is worth noting that the closest competence aim is that the student is expected to be able to *"*express and justify their own opinion on different topics" (UDIR, 2013b). This emphasizes an important part of this assessment criteria, as it does not only focus on whether the student can answer the question but also on how the student can reflect on the topic prompted by the question. The description given for a low level of achievement is that the student "can answer questions in a limited way", for a medium level of achievement it is "can answer questions and give some reasoning",

while a high level of achievement demands that the student can "reflect and give the reasoning for the answers". Thus, the examiners need to interpret what "*limited*" and "*reasoning*" means.

This, however, raises the same challenge as in the previous assessment criteria, as it is still a question of degree when it comes to how examiners should operationalize the verb *reasoning* and its associated adjectives. It is interesting to note that the value-loaded term "good" often used by the examiners is not present in the assessment criteria, and instead, the examiners have to interpret what qualifies as answering questions "in a limited way", and what the difference between "some reasoning" and "the reasoning" is.

The examiners rarely discuss what level they consider the students' achievement to be at. During ME1, when looking at how the student answered questions, the examiner noted that the student "was answering lots of questions, seeming reflective and open and kind of trying to look at it from different areas". The exact quality of the student's reflections is not discussed, which can at least in part be explained by how there is no adjective in front of the words that have to be interpreted. As stated by Johannessen (2018), the fact that the criteria have to be both general and yet possible to assess is difficult without a shared understanding of what is adequate, what is good and what is excellent performance. The examiners bring up the strong and weak sides of the students' presentation, but it is not stated how these aspects interact until a grade is agreed on at the end.

The discussion of how well the student reflected on the questions took place at the end of the examination. That was when the examiners were listing key aspects of the student's performance in terms of what they considered to be good. These discussions at the end of the examiner conversations generally followed a structure of listing up what the student did that was good, what they did that was less good, and finally what grade the examiners considered to be appropriate. The same arguments brought up during these discussions were also mostly the same as what the students were given as feedback alongside their grade. However, given that they mostly summed up arguments that had been used before they usually did not include any new details.

5.1.4 Non-criterion-based aspects of the students' performance

Lastly, it is worth pointing out that non-criterion-based aspects of the student's performance were only brought up while discussing assessment criteria related to content. This suggests that these aspects of a student's performance were brought up when the examiners felt that something important was missing from the criteria. What level of achievement the student is at is something the examiners themselves need to reach an understanding of, while also being mindful of their interpretations and possible misinterpretations (Sandvik, 2013). One approach the examiners used to reach a shared understanding was to look holistically at the aspect of the students' performance described by the criteria. Another approach was to focus on granular and observable portions of the student's performance, like what advanced words the student used, and then discuss them. This will be further elaborated on in section 5.3.

5.2 Holistic overview vs granular aspects

The results from the observation of the examination revealed that there were certain strategies that the examiners utilized when they were communicating their interpretation of the assessment criteria and how it applied to the student's presentation.

5.2.1 Holistic Overview

One strategy was to look at the presentation as a whole and see what level of achievement was appropriate for that particular assessment criterion based on this holistic understanding. An example of this, which will be discussed in the next paragraph, is how the teachers looked at fluency or communication.

When looking at fluency during the examination, the examiners do not go into detail on the topic of fluency, instead stating that the student is some degree of fluent. As noted by Bøhn (2019) and Luoma (2004), fluency is something teachers can interpret in different ways, from focusing on the students' speech rate and the number of times they hesitate to only look for particularly notable occurrences when the student hesitates. Or the teacher might not do this at all, and instead, base the grade on their intuitive understanding of how good the students' fluency is. While the individual examiner in this study might be aware of how they assess fluency and the strengths and weaknesses of their particular method, this is not communicated to their co-examiner during the examination. While the analysis showed that the examiner did

not disagree with the presence of fluency and agreed that the students' fluency can be described as good, it is not clear what each of them was looking at when they came to this conclusion.

When looking at communication, the examiners seemed to indicate that the student's tendency to speak faster when they were stressed made it harder for the student to communicate their ideas. They also stated that the grammatical errors in the students' use of prepositions and word choices did not negatively impact her communication and that she communicates well. As Iwashita et al. (2008) point out, problems with pronunciation can negatively affect the whole presentation as examiners cannot assess what they did not hear or understand. Aalandslid (2018) found that teachers focused on avoiding breakdowns in communication and that they viewed fluency and pronunciation as key in facilitating this. While it is noted that rushing negatively impacts communication and that grammatical errors do not necessarily impact their communication, it is not clear what the examiners define as good communication. It can, however, be inferred that communication is viewed as acceptable as long as there are no problems that negatively affect it. Both examples of looking at assessment criteria from a more holistic perspective belong to the language construct, perhaps because it was more difficult to find specific examples to discuss.

5.2.2 Granular aspects

Another strategy was to focus on a specific aspect of the assessment criteria, and then focus on that while discussing the criteria it is a part of. When the examiners looked at the students' vocabulary, they stated that the student's vocabulary is quite good and referenced specific words they consider to be examples of this high-level language. And when a student's vocabulary is found lacking, it is noted that this is because it is not wide enough, and they assign a score based on this.

When looking at vocabulary, the examiners looked at low-frequency words but did not discuss the high-frequency ones. Examiners focus on how the students provide evidence of the richness of their lexicons by looking at what low-frequency words they use, while they don't mention how the students utilize high-frequency words. Because these high-frequency words are common in speech, being able to use them appropriately is also a sign of advanced speaking skills, as noted by Luoma (2004). While it is possible it was not brought up because it did not influence the students' language to a noticeable degree, it is worth noting that this is an aspect

of vocabulary that was not discussed. Given that high-frequency words are used more often, it would have been easier to find evidence of them being present.

When the examiners discuss a student's grammar, they point out that the student has problems with subject-verb concord and that these problems persisted throughout his presentation. The examiners bring up specific sentences which the student pronounced with the wrong structure and point out that this is strange given the students' otherwise strong language. While both examiners agreed that the subject-verb concord mistakes were a weakness, this is not necessarily a universal interpretation. For example, from a lingua franca perspective, aspects like tense are viewed as important as they can impact communication, while aspects like subject-verb concord are viewed as less important (Bentsen, 2017). Because there is no national standard for whether this is the right interpretation or not, it is left up to the examiners themselves to decide. Still, specific examples like the ones brought up by the examiners made it clearer what exactly they were focused on, and what they viewed as the weaknesses which affected the students' presentation the most.

With regard to pronunciation, the examiners viewed Norwegian-accented pronunciation as a weakness. During their discussions, the examiner pointed out that the student pronounced keywords like NATO and satellite with a Norwegian accent. This was viewed as a weakness, as these are important terms. This is in keeping with Aalandslid (2018)'s findings, where she noted that the teachers viewed students speaking English with a Norwegian accent as a threat to intelligibility. While the teachers in her study stated they did not believe that students had to have a native-like pronunciation, they did not want it to be impacted by Norwegian either. As Luoma (2004) points out, achieving native-like pronunciation is difficult, and most students would receive a lower grade if they were assessed according to this standard.

What accent the students use can influence how the examiners assess them, based on what accents the examiner view as the goal In Bøhn (2016)'s doctoral thesis, one of the disagreements between teachers was whether a native-like pronunciation should be valued higher. While the examiners in my study did not discuss their views on accents, it might have impacted their views on pronunciation. Viewing the students' L1 as a weakness when it influences the student's English is not the only possible interpretation. As Iannuzzi and Rindal (2018) point out, according to a Lingua Franca perspective, the student's native tongue is an advantage as long as it improves the communication. This highlights the importance of

examiners being aware of their language ideology and their biases, so they can keep it in mind during assessment.

5.2.3 Content

Looking at the tasks holistically and more granularly is not unique to the language construct. The same processes take place when the examiners look at criteria related to content. When looking at the students' ability to reflect on the questions they were asked, the examiners point out when a student can answer a lot of questions while at the same time being reflective and open about their answers. Meanwhile, when the student can answer questions but doesn't reflect on them, this is also pointed out.

When the examiners looked at the student's understanding of the topic they were discussing, they found that the student could talk well about the cold war in general. However, the student was unable to say anything about the differences between communism and capitalism, which the examiners noted was a clear weakness. While discussing another student's conversation about native people, the examiners noted that her knowledge about the particular topic she had been told to prepare to talk about as well as her general knowledge was weak. They also focused on very specific details that the student did not know, like when an examiner noted that the student had forgotten the name of the neighborhood in which the book takes place.

5.2.4 The impact of different levels of detail

The level of detail that the examiners decide to look at during an oral examination will affect their view of the student. It is difficult to look at both specific details and focus on the overall image within the constraints of an examination situation. To make sure that they have the same understanding of an assessment criterion, the examiners would have to separate them into smaller pieces and then look at these pieces together to see how the students score on each of them. While this has the benefit of making the student's actual level of achievement more easily visible to the examiners, it can also cause the overall competence aims to become blurred, which might cause further confusion (Hartberg, Dobson & Gran, 2012). Additionally, this process takes time and it also makes it easier for the examiners to get lost in the details rather than keeping a holistic overview of how well the student performed in general.

This overview will inevitably be influenced by subjective opinion, as each examiner has to weigh the different parts of the student's performance that they observed and then assign a score they consider appropriate for the sum of the parts. Given the time constraints of an oral examination, as well as the limitations on the number of factors that an examiner can notice at the same time, some parts of the student's performance will receive less focus than other parts. And given that each examination is done only one time and discussed immediately afterward with no recording available, anything the examiners missed the first time cannot be brought up. The quality of the evidence available to each examiner varies, which can influence the reliability and validity of the assessment situation as it affects what the examiners base their interpretations on.

The balancing act between having a clear overview and a focus on particular details will influence the reliability of the examinations. In addition to having to choose what to focus on during the discussion afterward, the examiner will also have to choose what to focus on during the presentation itself. A benefit of the system with two examiners, however, is that it makes it more likely that a particular aspect of the presentation is brought up, as two people are likely to focus on different parts of the presentation. One way to help with the balancing of overview and details is to standardize what the examiners are to look for, but this has its own set of advantages and drawbacks, as will be discussed in the next section.

5.3 Autonomy vs Standardization

English teachers in Norway must follow the national English subject curriculum (UDIR, 2013b), which is made by the Norwegian Directorate for Education and Training. This is done to ensure that their students fulfill the competence aims laid out in the curriculum. At the same time, the examination tasks which the students are answering and the assessment criteria by which these tasks are assessed are made locally. Balancing between centralized standardization of the curriculum and the local development of the assessment situations have important implications for examinations, given their high stakes.

One of the things that were discovered during the analysis of the examinations was that the examiners brought up features of the student's performance during the examination which were not referred to in the assessment criteria they had in front of them while assessing. As noted by Cummin (2013), such sources of rater variance can have outsized impacts on the students

affected by it, because of the importance of grades in today's society. It is therefore important to limit the impact it has on the student's grades. The non-criterion-based features that were brought up during the assessment situation were sorted into three categories during the analysis: *Comparison with other students*, *outstanding quality*, and *teachers' intuition*.

5.3.1 Comparison with other students

Examiners utilized comparisons by looking at how well a previous student managed to fulfill an assessment criterion, and then their performance was compared to how the student in the current examination fulfilled the criteria. It is worth noting again that, as the Norwegian system is criteria-based, such comparisons should not take place. This was done either to note that the students' analysis was not as good as what the student who was at the highest level had been able to do earlier, or that the student's general understanding and knowledge of the topics she discussed was weaker than what the examiner had seen previously.

The comparison was utilized by the examiners to help define what level of understanding the student had, by using more concrete examples from earlier. This is similar to what was found in Ang-aw and Goh (2011)'s study, where the examiners compared the students' performance to what others had done before them as they graded the performance. While comparing one student to another might lead to one student's performance influencing the examiner's view of another student, the examiners kept their comparisons narrowly focused on specific assessment criteria and did not compare one student's whole presentation with another student's whole presentation. The role of the comparison was limited to specific criteria, and they did not compare how one student did overall compared to another student.

5.3.2 Outstanding quality

The outstanding quality was brought up by the examiners when they focused on the student's creativity, unique thoughts, and risk-taking and when the examiners noted that six-level students needed to display some kind of outstanding quality that the current presentation did not have. While this is not mentioned in the assessment criteria, this is something that the examiner viewed as an important aspect of the student's performance. It is worth noting that both the presence and the lack of outstanding quality were brought up by the examiners, and thus it could either help the student get a better grade or be one of the things that kept them from a higher grade.

5.3.3 Teachers' intuition

The teacher's intuition was brought up once when the examiner states that she did not feel that the student's performance is at a six level. This statement fits well with Ang-Aw and Goh (2011)'s observation that ambiguity in the assessment criteria could cause the examiners to depend more on what their feelings tell them regarding whether a specific grade is right for the student or not. However, this intuition was not brought up in isolation and was subsequently backed up by references to what the examiner felt the presentation lacked.

5.3.4 Assessment criteria and lack of guidelines

The assessment criteria that the examiners were using to assess the students were of a generalized nature as they included assessment criteria irrelevant to the assessment situation. The fact that the criteria were not adapted to the test might have caused more subjectivity for the teachers, as the criteria they were to interpret were not always relevant to the task at hand. While directly referencing the criteria might have helped the examiners build a shared understanding of what they were to assess, the vagueness inherent in the criteria might have limited the effectiveness of this. It is also worth noting that there were no cases of disputes between the examiners, and they generally agreed on the student's level of achievement when criteria were brought up. Additionally, they also generally agreed on the grades given to the students at the end of the examination. This is similar to Bøhn (2016)'s study, where most agreed on each criterion and the final grade, and the disagreement was mostly regarding whether one criterion was more important than any other.

As the assessment criteria are made at the local level with no national standardized guidelines, it is left up to the individual teacher to define what each aspect of the assessment criteria means, and subsequently what they should look for during the assessment. As the words and phrases used in the assessment criteria are somewhat vague, this causes the teachers' subjective opinions to play a large role in how they assess a student's level of competence. Because of this, previous research has suggested a need for national guidelines that teachers can use for oral examinations (Bøhn, 2016; Yildiz, 2011), which my thesis also supports. The vagueness of the assessment criteria causes the teachers to look at other aspects of the students' performance, which will negatively influence reliability as it is not done by everyone.

At the same time, there are some benefits to subjectivity. As stated by Taylor and Galaczi (2011), a standardized rating scale can end up leading to the introduction of assessment criteria which are too narrowly defined to adequately capture the complexity inherent in the criteria which are tested in the assessment. While it might be relatively easy to define clear guidelines for something concrete like the presence or lack of subject-verb concord, this is more difficult to do for large and abstract concepts like communication. This might cause the examiners to lose sight of the student's overall level of fluency if they are too focused on how many gaps there are in the student's speech. In such cases, the examiners will have to depend on their intuition as it is impractical to have

More standardized assessment criteria may impact what students are taught. If what is to be assessed are factors that are easy to document and interpret, then this can lead to a lack of focus on more advanced intellectual activities which by their nature would be difficult to reliably document (Moss, 1994). This can then lead to more cases of educators teaching the test rather than maintaining a holistic view of the student and their overall development. While quality assurances need to be stringent for something as important as examinations (Harlen, 2012), they should not come at too great a cost to the student's general progression in the subject.

Additionally, bringing up non-criterion-based features of the student's performance is not solely a negative thing. Given the abstract nature of assessment criteria, references to something both of the examiners had seen previously might help them ground their interpretations of the student's performance in a shared understanding of a previous student's performance. This, in turn, can help make the vague criteria more concrete by referencing a specific situation as a baseline of understanding. A key aspect of any test, as noted by Black and William (2012), is to ensure that the variations in test scores are caused by differences between students that are relevant, and not caused by irrelevant factors like who does the grading or what part of the curriculum were chosen for the test. If what is looked at is actually relevant yet not one of the assessment criteria, it can be argued that discussing it is acceptable.

Finally, it is worth noting that comparing one student to another is not inherently bad, as long as it is used to establish a shared understanding. Looking for outstanding quality can be sensible as long as it is done fairly and reliably by all examiners. And relying on one's intuition can be beneficial when faced with assessment criteria that make it difficult to put one's understanding of the student's performance into words, as long as the examiner does not rely on too much.

However, it is worth reiterating that in a criteria-based system like the Norwegian one, this should not occur. And as long as some examiners bring up these aspects and others do not, it will negatively affect the reliability and validity of the assessment situation, which is why standardization is important.

6.0 Conclusion

The study in this thesis has investigated how examiners operationalize the assessment criteria they are given during an oral examination in 10th grade, by observing two examiners during their assessments of five students. During their discussions, the examiners reviewed the students' performance during the examination and discussed aspects that they considered particularly relevant. The analysis of the audio-recorded discussions revealed that the examiners focused on both content and language. It also showed that they generally agreed on what narrow aspects of the student's performance were important during their conversation after the examination, when they were discussing criteria like vocabulary or fluency. Finally, it is worth noting that aspects of the student's performance that were not represented in the assessment criteria were brought up

In this section, I will conclude by summarizing the empirical contribution of the study and presenting the implications that this thesis has for assessment in oral L2 English examination situations (6.1). Then, the limitations of the thesis will be reviewed (6.2) and further research will be suggested (6.3).

6.1 Summary of findings

The present thesis found that the examiners utilized various strategies to operationalize the assessment criteria. Some criteria, like the ones related to vocabulary, were discussed extensively by the examiners. They pointed out specific words the student had used that were viewed as impressive and commented on the richness of their internal lexicon. Other criteria, like the one related to fluency, received less attention from the examiners. The comments regarding fluency were limited to ascribing a level of achievement to the students, like describing it as "good", or in one case simply noting that fluency was present in their response. It is also worth noting that the examiners' individual preferences with regard to criteria like accents or vocabulary were not discussed nor brought up. This will be discussed further below, as it might impact their understanding of what constitutes "good" or "bad" performance.

While there were fewer assessment criteria for *content* compared to *language*, they had about the same number of mentions in total. Some criteria were referenced more than others, with the criteria relating to the student's understanding of the topic receiving the most attention.

Compared to Bøhn (2016)'s findings, where he found that upper secondary teachers focused more on either content or language depending on their students, this balance is noteworthy. It is, however, not clear if this is because of any inherent quality of lower secondary school, or merely caused by the views of the teachers involved.

Lastly, in addition to what was mentioned in the assessment criteria, the examiners utilized other aspects of the student's performance as topics of discussion. One such aspect was the examiner pointing out that the student's presentation either had or lacked a particularly outstanding aspect expected of those receiving the highest grade. Another such aspect was comparing how well one student had answered an assessment criterion, compared to how the current student had answered. These non-criterion-based assessment features seemed to be used by the examiners to help them decide how well a student did, by looking at a concrete example they both had a shared understanding of.

6.2 Implications for Assessment

While the teachers' perceptions of the different aspects of the student's performance showed a considerable degree of agreement, it is worth noting that their discussions of the assessment criteria did not include the actual adjectives used in the criteria themselves. Instead, the student's performance was described with value-loaded language like "good" or "bad" or referenced only when the examiners remarked that a particular criterion was present without providing further comment. It is reasonable to believe that this is caused by the relative vagueness of the criteria and the adjectives it uses to describe the different levels of the student's performance. It was noted by Bøhn (2016) and Yildiz (2011) that vagueness in criteria causes the subjective opinion of the examiners to matter more. Thus, examiners must be aware of their own biases during an assessment situation.

Ideally, the guidelines given to the examiners could offer explanations as well as definitions of the different aspects of the assessment criteria. There seems to be a need to define what fluency is and what it refers to, as both Bøhn (2016) and Luoma (2004) show that examiners can interpret and thus operationalize this concept in very different ways. The same can be said for vocabulary, where the focus seems to be mostly on the richness of the students' inner lexicon, although scholars argue that knowledge of and the right use of high-frequency words is equally important (Luoma, 2004). Detailed guidelines might help remind examiners of this. It might

also help examiners ascertain what students are supposed to know at their current stage in education, which might help make the content construct clearer to them. Audio-recordings similar to those used as data material for this study might be useful for developing assessment practices, as they can be utilized as practical examples for teachers or teacher students to discuss during their training. This would be similar in role to the model texts that are used during rater training for written examinations.

The interpretations which the examiners make of the assessment criteria are likely to be influenced by ideas of correctness and beliefs about what constitutes "good" language. More rater training related to language learning paradigms might raise awareness about such ideas and beliefs. Teachers are given little guidance on what is considered the correct paradigm, as this is left up to the individual teacher. This means that they could benefit from increased awareness regarding how such things could affect their judgment, particularly how conflicting subconscious beliefs might lead to differing opinions. When the examiner brought up the students' use of Norwegian words as well as a Norwegian pronunciation of words during the examination, they noted that this was a weakness. While this is a valid interpretation if one views language as separate entities that should not be mixed, it is equally valid to view the use of one's L1 as a strength. If one followed the lingua franca core, as laid out by Iannuzzi and Rindal (2018) as one possible model for intelligibility, the use of the student's native tongue would have helped the communication move along.

While the examiners did not discuss the role of accent, language ideology can influence this aspect as well. If an examiner views Norwegian-accented English as a weakness, this might very well cause them to view the students' language competence as weaker than one with a General American or Received Pronunciation accent. As found by Luoma (2004), most learners would receive lower grades if assessed by the standards of native speakers. There are therefore reasons to believe that teachers interpret aspects of the student's competence differently, like Bøhn (2016)'s finding that examiners focus on different aspects of the assessment criteria. Because the presence or lack of such features might impact the examiners' view of the students' competence, it is important to be aware of these factors, as it can make it easier to limit the impact of one's biases.

Additionally, the fact that the examiners brought up factors not referred to in the assessment criteria is interesting. One of the examiners noted that she did not feel that the student's

performance warranted the highest grade. Ang-Aw and Goh (2011) note that ambiguity in the assessment criteria might cause teachers to rely more on their intuition during assessments. While it can be useful, this intuition will also vary from examiner to examiner, and should thus be used sparingly or in combination with clear evidence.

It is important to be aware of the balance between teacher autonomy and standardization. Taylor and Galaczi (2011) note that too narrowly defined assessment criteria might make it more difficult for examiners to focus on the larger picture as well. As Harlen (2012) points out, stringent quality controls should not come at the cost of students' progression, which might be the consequence of too standardized testing. This means it might be counter-productive to eliminate too much of the uncertainty as it might create another source of variation. While intuition and concrete references to earlier students are useful tools for the examiners, it is important that they are aware of how this might influence their assessment of the current student as well as those that follow. The implications of this study is that examiners need more awareness of their own beliefs and attitudes, and how this affect the validity and reliablity of assessment.

6.3 Limitations on the Thesis' Relevance

This thesis was based on collected material from a school at a time where the previous curriculum (LK06) was still used. This earlier curriculum differs from the current curriculum (LK20) in ways relevant to the purpose of this thesis. One example of this is that a criterion about discussing how people live, which used to refer explicitly to "Great Britain, USA and other English-speaking countries and Norway" now simply refers to "the English-speaking world". The new formulation removes the previous preference for nativeness, and thus gives the individual teacher more freedom in what they want to teach their class. However, this autonomy also means that the actual content taught in each classroom might differ more than previously. This makes it more important that examiners communicate with each other to ensure they have a shared understanding of what the student is expected to know.

Another key aspect of the new national curriculum is the focus on in-depth learning, which involves the gradual development of knowledge across different subjects, as well as how to utilize different methods and how to reflect on one's own learning (ref). This focus on metacognition could influence particularly the content construct. Bøhn (2016) noted that the

only discrepancy in his study was related to metacognition, where one rater denied that it was to be tested even though it was referenced in the curriculum. It is thus likely that the examiners' understanding of the content construct as a consequence of the implications of LK20, which will make it interesting to see what strategies examiners will employ to operationalize it.

6.4 Suggestions for further research

As mentioned previously, the new curriculum will impact future examinations. Because of this, it would be interesting to see future research into oral examinations, and how they are impacted by the implementation of LK20 (UDIR, 2020).

While there were no situations where significantly divergent assessment occurred in my study, the possibility were there. The vagueness of the assessment criteria, the use of factors not related to said criteria, and the potential influence of language ideology mean that many factors could lead to this. One example of a study looking into this would be Sandlund and Sundqvist (2016). They looked at an assessment situation with two interlocutors talking about the same topic with the examiner present. This study revealed what factors might impact a situation where two people are talking together when their conversation is what is meant to be assessed. This calls for further research into the difference between examiners present in the situation and those external to it, and how such factors might impact the reliability and validity of the assessment. It would be interesting to see if they understand the assessment criteria in different ways.

Finally, a longitudinal study with several pairs of examiners would be interesting, as that would enable the study to review multiple different viewpoints and look at how their strategies evolve and change over time. My study focused on two raters during one examination, which meant that the behavior which could be observed was limited. While looking at several examiners over time would be logistically difficult because of the randomized pairing of examiners during examinations, it would provide a unique insight into the examiners' thought processes. It would be particularly interesting to discover their perspectives on their own language ideology and see if any patterns can be found regarding how that influences their discussions and what they focus on.

References

Ang-Aw, H. T. & Goh, C. C. M. (2011). *Understanding discrepancies in rater judgement on national level oral examination tasks.* RELC, 42(1), 31-51. doi: 10.1177/0033688210390226

Aalandslid, E. (2018) *Fluency and stuff: Perceptions of oral competence among teachers and students in vg1*. (Master thesis). Oslo: University of Oslo. Retrieved from https://www.duo.uio.no/handle/10852/63514?show=full

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Anderson-Bakken, E & Dalland, C. (2021). *Metoder i klasseromsforskning : forskningsdesign, datainnsamling og analyse*. Oslo: Universitetsforlaget

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bagarić, V., & Djigunović J.M. (2007). *Defining communicative competence*. Metodika, 8(1), 94-103.

Bakke, M. H. (2010). *Teaching reading in EFL-instruction*. (Master's thesis). Oslo: University of Oslo

Befring, E. (2015). *Forskningsmetoder i utdanningsvitenskap.* Oslo: Cappelen Damm Akademisk.

Bentsen, L. (2017), *To teach or not to teach grammar? – Teachers' approaches to grammar teaching in lower secondary school.* (Master's thesis) Oslo: University of Oslo.

Black, P., & Wiliam, D. (2012). *The reliability of assessments.* In J. Gardner (Ed.), Assessment and Learning (pp. 243-263). London: Sage.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *The Taxonomy of educational objectives, handbook I: The Cognitive domain*. New York: David McKay Co., Inc.

Brevik, L. M. (2015). *How teachers teach and readers read. Developing reading comprehension in English in Norwegian upper secondary school*. (Doctoral dissertation). University of Oslo, Oslo.

Brevik, L. M. & Doetjes, G. (2020). *Tospråklig opplæring på fagenes premisser.* Rapport fra Evaluering av tospråklig opplæring i skolen (ETOS-prosjektet). University of Oslo, Oslo.

Brown, J. D. (1996) *Testing in Language programs.* Prentice Hall, inc.

Bøhn, H. (2016). *What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway.* (Doctoral thesis) Oslo: The University of Oslo. Retrieved from: https://www.duo.uio.no/handle/10852/53229

Canale, M., & Swain, M. (1980). *Theoretical bases of communicative approaches to second language teaching and testing*. Applied Linguistics,1, 1-47. https://dx.doi.org/10.1093/applin/I.1.1

Chalhoub-Deville, M. (2003). *Second language interaction: Current perspectives and future trends*. Language Testing, 20(4), 369-383.

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Council of Europe, Language Policy Unit North, 2004

Creswell, J. W. (2014). *Research design: Qualitative, Quantitative and Mixed Methods Approaches* (4th Ed.). Los Angeles: SAGE.

Creswell, J. W. (2007). *Qualitative inquiry & research design: choosing among five Approaches* (2nd ed.). Thousand Oaks: SAGE Publications.

Creswell, J., & Miller, D. (2000). *Determining Validity in Qualitative Inquiry. Theory Into Practice*, 39(3), pp. 124-130.

Cumming, A. (2013). Validation in langauge assessments. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 6006-6015). Chichester, UK: Blackwell.Sandlund and Sundqvist, 2016

Douglas, D. (1994). Quantity and Quality in Speaking Test Performance. *Language Testing*, 11(2), 125-144.

Eriksen, H. (2018). *Læringsfremmende vurderingspraksis – med blikk på norskfaget i to skoler.* (Doctoral thesis) Oslo: The University of Oslo. Retrieved from: https://www.duo.uio.no/handle/10852/61824

Eurydice. (2008). *Levels of Autonomy and Responsibilities of Teachers in Europe*. Retrieved from http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/094EN

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. Oxford: Routledge.

Opplæringsloven. (2020) *Forskrifta til opplæringsloven*. (FOR-2006-06-23-724). Lovdata. Retrieved from https://lovdata.no/dokument/SF/forskrift/2006-06-23-724?q=Oppl%C3%A6ringsloven

Gleiss, M. & Sæther, E. (2021). *Forskningsmetode for lærerstudenter. Å utvikle ny kunnskap i forskning og praksis. Oslo:* Cappelen Damm Akademisk

Green, A. (2014). *Exploring Language Assessment and Testing*. Language in Action. New York: Routledge.

Hallgren, K. A. (2012). *Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial*. Tutor Quant Methods Psychol, 8(1), 23-34. Bryman, 2016

Harlen, W. (2012). *On the relationship between assessment for formative and summative purposes.* In J. Gardner (Ed.), Assessment and learning (2nd ed., pp. 87-101). London: Sage.

Hartberg, E.W., Dobson, S. & Gran. L. (2012). *Feedback i skolen*. Oslo: Gyldendal Akademisk.

Hymes, D. H. (1972). *On Communicative Competence*. In J. B. Pride, & J. Holmes, (Eds.), Sociolinguistics (pp. 269-293). Baltimore, USA: Penguin Education, Penguin Books Ltd.

Iannuzzi, M. E. & Rindal, U. (2018) Uttaleundervisning i verdensspråket engelsk. In Bedre skole, 1. Retrieved from: https://www.utdanningsnytt.no/bedre-skole/debatt/2018/januar/uttaleundervisning-i-verdenssprake-engelsk/

Iwashita, N., Brown, A., McNamara, T. & O'Hagan, S. (2008). *Assessed Levels of Second Language Speaking Proficiency: How Distinct?, Applied Linguistics*, 29(1), 24–49. https://dx.doi.org/10.1093/applin/amm017

Jenkins, S. & Parra, I. (2003). *Multiple Layers of Meaning in an Oral Proficiency Test: The Complementary Roles of Nonverbal, Paralinguistic, and Verbal Behaviors* in Assessment Decisions. The Modern Language Journal, 87(1), 90-107. https://doi.org/10.1111/1540-4781.00180

Johannessen, S. L. (2018). *Oral assessment in the English subject. Teachers' understandings of what to assess.* (Master's thesis, Inland University). Retrieved from https://brage.inn.no/inn-xmlui/handle/11250/2560214

Johannessen, A., Tufte, P.A., Kristoffersen, L. (2005) *Introduksjon til samfunnsvitenskapelig metode*. Oslo: Abstrakt forlag

Johnson, B. R. (2013). Validity of Research Results in Quantitative, Qualitative and Mixed Research. In B. R. Johnson & L. Christensen (Eds.), *Educational Research: Quantitative, Qualitative and Mixed Approaches* (pp. 277-316). Los Angeles: Sage.

Johnson, B. R., & Christensen, L. (2017). *Educational research: Quantitative, qualitative, and mixed approaches* (Sixth ed.). Thousand Oaks, California: SAGE

Kim, H. J. (2015). *A qualitative analysis of rater behavior on an L2 speaking assessment*. Language Assessment Quarterly, 12(3), 239-261. doi:10.1080/15434303.2015.1049353.

Klette, K., & Blikkstad-Balas, M. (2017) *Observation manuals as lenses to classroom teaching: Pitfalls and possibilities.* European Educational Research Journal. Retrieved from https://journals.sagepub.com/doi/epub/10.1177/1474904117703228

Kleven, T. A., Hjardemaal. F. & Tveit, K. (2014). *Innføring i pedagogisk forskningsmetode: en hjelp til kritisk tolkning og vurdering* (2nd ed.). Bergen: Fagbokforlaget

Kvale, S. & Brinkmann, S. (2015). *Det kvalitative forskningsintervju*. Oslo: Gyldendal

Larsson, S. (2009). *A pluralist view of generalization in qualitative research*. International Journal of Research & Method in Education, 32(1), 25-38. doi:10.1080/17437270902759931

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369-377

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.

Maxwell, J. A. (2013). *Qualitative Research Design; An Interactive Approach* (3rd ed.). Los Angeles: Sage

McNamara, T. (2000). *Language Testing*. Oxford University Press.

Met, M. (1998). Curriculum decision-making in content-based language teaching. In J. Cenoz & F. Genesee (Eds.), *Beyond bilingualism: Multilingualism and multilingual education* (pp. 35-63). Philidelphia, PA Multilingual Matters.

Mitchell, J. C. (1983). Case and situation analysis. *Sociological Review*, 31(2), 187-211.

Moss, P. A. (1994). *Can there be validity without reliability?* Educational Researcher, 23(2), 5-12. doi:10.3102/0013189x023002005.

Newton, P. E. (2012). *Clarifying the consensus definition of validity*. Measurement: Interdisciplinary Research and Perspectives, 10(1-2), 1-29. doi:10.1080/15366367.2012.669666.

Nusche, D., Earl, L., Maxwell, W.,& Shewbridge, C. (2011). *OECD Reviews of Evaluation and Assessment in Education: Norway*. Retrieved from https://www.oecd.org/norway/48632032.pdf

Orr, M. (2002). *The FCE speaking test: Using Rater Reports to Help Interpret Test Scores*. System, 30, 143-154. Retrieved from https://ac-els-cdn-

com.ezproxy.inn.no/S0346251X02000027/1-s2.0-S0346251X02000027-main.pdf?_tid=c16cdc5c-7e82-455c-bb29777413a51839&acdnat=1555501455_819e4873fc71ff75b92331ed6bd77ede

Ragin, C. (1994). *Construction Social Research.* Thousand Oaks, California: Pine Forge Press

Read, J. (2000) *Assessing Vocabulary*. Cambridge: Cambridge University Press

Rindal, U. (2014): What is English? in *Acta Didactica Norge*. Art 14, 8(2)

Rindal, U. (2015). Who owns English in Norway? In A. Linn, N. Bermel & G. Ferguson (eds): *Attitudes towards English in Europe*. (pp. 241-270). Berlin/Boston: Walter de Gruyter

Sandvik, L.V. (2013) Perspektiver på individuell vurdering i skolen. In Sandvik, L. V., Buland, T. (Eds.), *Vurdering i skolen. Operasjonaliseringer og praksiser*. Retrieved from https://www.researchgate.net/publication/277140787_Vurdering_i_skolen_Operasjonaliseringer_og_praksiser_Delrapport_2_fra_prosjektet_Forskning_pa_individuell_vurdering_i_skolen_FIVIS

Silverman, D. (2011). *Designing a research project: Interpreting Qualitative Data*. (4th ed.). Thousand Oaks: Sage

Simensen, A.M. (1998). *Teaching a Foreign Language – Principals and Procedures*. Bergen: Fagbokforlaget

Snow, M. A., & Katz, A. M. (2014). Assessing language and content. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, pp. 230-247). Chichester, UK: Wiley Blackwell.

Stobart, G. (2012). Validity in Formative Assessment. In J. Gardner (Ed.), *Assessment and Learning* (pp. 233-242). Sage: Los Angeles.

Taylor, L., & Galaczi, E. (2011). Scoring Validity. In L. Taylor (Ed.), *Examining Speaking: Research and practice in assessing second language speaking* (Vol. 30, pp. 171-233). Cambridge: Cambridge University Press.

The Norwegian Universities and Colleges Admission Service (2020). *Ordinær kvote*. Retrieved from http://www.samordnaopptak.no/info/opptak/opptakskvoter/ordinerkvote.html

UDIR [LK06] (2013b). *English subject curriculum* (ENG1-03). Retrieved from: https://www.udir.no/kl06/eng1-03/

UDIR [LK20] (2020). *English subject curriculum* (ENG01-04). Retrieved from: https://www.udir.no/lk20/eng01-04/

KD (2013a) *Muntlig eksamen*. Oslo: Ministry of Education and Research. Retrieved from: http://www.udir.no/Vurdering/Eksamen/Muntlig-eksamen/?WT.ac=muntlig&boks=pavirker

KD (2021) Indikatorveiledning – Ståstedsanalysen og tilstandsrapporten for grunnskolen. Oslo: Ministry of Education and Research. Retrieved from: https://www.udir.no/kvalitet-og-kompetanse/om-statistikken--tilstandsrapporten-for-grunnskolen/karakterer-og-grunnskolepoeng/

Wickström, G., & Bendix, T. (2000). *The "Hawthorne effect" - what did the original Hawthorne studies actually show?* Scandinavian Journal of Work, Environment & Health, 26(4), pp. 363-367.

Yildiz, L.M. (2011). *English VG1 level oral examinations: How are they designed, conducted and assessed?* (Master thesis). Oslo: University of Oslo. Retrieved from: https://www.duo.uio.no/handle/10852/32421

Yin, R. K. (2016). *Qualitative research from start to finish* (2nd ed.). New York: The Guilford Press.

Appendix

Appendix 1: Assessment criteria (Turned into codes)

Low	**Medium**	**High**	**Codes**
Presents a topic by reading from a script	Uses a script in a free way (…) talks about (texts) in a free way based on some notes.	Speaks freely	*Script*
Switches a little between English and Norwegian pronunciation	Has mostly correct pronunciation	Has a very good pronunciation	*Pronunciation*
Knows some English words and expressions and uses them. Communicates a message through the use of simple words and expressions	Knows many English words and expressions and uses these to describe various subjects	Has a large vocabulary that is advanced, varied, and descriptive and can use it in a purposeful way in different subjects with varied content	*Vocabulary*
Often appearances of Norwegian words	The language has some appearances of variation, nuance, and idioms	Uses the language precisely, nuanced, and varied with appearances of idioms/English expressions	*Idioms*
Has some familiarity with grammar	Has good grammar	Has correct grammar	*Grammar*
Puts together words so it makes sense and forms understandable	Expresses themselves with some fluency, precision, and	Expresses themselves precisely. Utilizes a language	*Fluency and cohesion*

sentences	cohesion. Presents timely themes with logical cohesion	with good fluency and cohesion	
Reads and communicates an understanding of short non-fiction texts with a low difficulty level	Reads and communicates an understanding of non-fiction texts with a medium difficulty level	Reads texts at all levels and reflects over these. Has very good reading- and language understanding, and has a good understanding of English expressions which are specific for the different subjects	*Reading comprehension*
Reads simple texts in different genres and about different subjects and retells these.	Reads different texts and retells these.	Compares, talks about, evaluates, and presents several varied and timely topics in a purposeful way	*Reading, evaluation, and comparison of texts*
Talks about the topic, with some signs of repetition from memory	Have obtained an understanding of the topic, and mentions causes and consequences	Have obtained a good understanding of the topic, and reflect on the causes and consequences	*Understanding of topic*
Can answer questions in a limited way	Can answer questions and give some reasoning	Can reflect and give the reasoning for the answers	*Reflection on questions*
Uses some content from sources	Uses content from varied sources and documents this in accordance with rules	Uses content from varied sources in an independent and critical manner, and documents this in accordance with rules	*Sources*

Appendix 2: Assessment criteria (Original Norwegian version)

Kjennetegn på måloppnåelse		
Lav	Middels	Høy
Presenterer et tema ved å lese opp fra manus	Bruker manus på en ledig måte	Snakker fritt
Veksler litt mellom norsk og engelsk uttale	Har stort sett riktig uttale	Har meget god uttale
Kjenner noen engelske ord og uttrykk og bruker noen dem	Kjenner til mange engelske ord og uttrykk og bruker disse til å beskrive ulike emner	Har et stort ordforåd som er avansert, variert, og beskrivende og kan bruke dette på en hensiktsmessig måte i ulike emner med variert innhold
Ofte innslag av norske ord	Språket har innslag av variasjon, nyanser og idiomer/engelske uttrykk	Bruker språket presist, nyansert og variert med innslag av idiomer/engelske uttrykk
Har litt kjennskap til grammatikk	Har god grammatikk	Har korrekt grammatikk
Setter sammen ord slik at det gir mening og former forståelige setninger	Uttrykker seg med noe presisjon, flyt og sammenheng	Uttrykker seg presist og anvender et språk med god flyt og sammenheng
Leser og formidler forståelse av korte faglige tekster med lav vanskegrad	Leser og formidler forståelse av faglige tekster med middels vanskegrad	Leser tekster på alle nivåer og reflekterer over disse. Har meget god lese- og språkforståelse og kjenner godt til engelske uttrykk som er spesifikt for de ulike emnene
Leser enkle tekster I forskjellig sjangere og om ulike tema og gjenforteller disse. Formidler et budskap med bruk av enkle ord og uttrykk	Leser forskjellige tekster og gjenforteller disse. Presenterer aktuelle temaer med logisk sammenheng og samtaler om dem på en ledig måte ut fra noen stikkord	Sammenlikner, samtaler om, vurderer og presenterer flere ulike aktuelle temaer på en hensiktsmessig måte
Forteller om tema, bærer noe preg av oppramsing	Har satt seg inn I temaet og nevner årsaker og konsekvenser	Har satt seg godt inn i tema og reflekterer over årsaker og konsekvenser
Kan svare enkelt på spørsmål	Kan svare med noe begrunnelse på spørsmål	Kan svare reflektert og begrunner svarene

Bruker noe innhold fra kilder	Bruker innhold fra ulike kilder og dokumentere disse I henhold til regler	Bruker innhold fra ulike kilder på en selvstendig og kritisk måte og dokumenterer disse I henhold til regler